WILEY

BIOMACROMOLECULES

Introduction to Structure, Function and Informatics

C. STAN TSAI, Ph.D.

VISIT…

LANZAROTE
Caliente.com

BIOMACROMOLECULES

BICENTENNIAL
1807
WILEY
2007
BICENTENNIAL

THE WILEY BICENTENNIAL–KNOWLEDGE FOR GENERATIONS

Each generation has its unique needs and aspirations. When Charles Wiley first opened his small printing shop in lower Manhattan in 1807, it was a generation of boundless potential searching for an identity. And we were there, helping to define a new American literary tradition. Over half a century later, in the midst of the Second Industrial Revolution, it was a generation focused on building the future. Once again, we were there, supplying the critical scientific, technical, and engineering knowledge that helped frame the world. Throughout the 20th Century, and into the new millennium, nations began to reach out beyond their own borders and a new international community was born. Wiley was there, expanding its operations around the world to enable a global exchange of ideas, opinions, and know-how.

For 200 years, Wiley has been an integral part of each generation's journey, enabling the flow of information and understanding necessary to meet their needs and fulfill their aspirations. Today, bold new technologies are changing the way we live and learn. Wiley will be there, providing you the must-have knowledge you need to imagine new worlds, new possibilities, and new opportunities.

Generations come and go, but you can always count on Wiley to provide you the knowledge you need, when and where you need it!

William J. Pesce
WILLIAM J. PESCE
PRESIDENT AND CHIEF EXECUTIVE OFFICER

Peter Booth Wiley
PETER BOOTH WILEY
CHAIRMAN OF THE BOARD

BIOMACROMOLECULES

Introduction to Structure, Function and Informatics

C. STAN TSAI

Department of Chemistry, Carleton University

WILEY-LISS

A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2007 by John Wiley & Sons, Inc. All rights reserved

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at http://www.wiley.com/go/permission.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:
Tsai, C. Stan.
Biomacromolecules : introduction to structure, function, and informatics / C. Stan Tsai.
p. cm.
Includes bibliographical references and index.
ISBN-13: 978-0-471-71397-5
ISBN-10: 0-471-71397-X (cloth)
1. Macromolecules. 2. Biomolecules. I. Title.

QP801.P64T73 2006
572′.33–dc22
2006040639

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

CONTENTS

Preface xiii
Abbreviations in Repetitive Use xvii

CHAPTER 1 *INTRODUCTION* 1

1.1 Prelude 1
1.2 Covalent Bonds 4
1.3 Noncovalent Interactions 5
  1.3.1 Electrostatic Interaction 6
  1.3.2 Van der Waals Interaction 6
  1.3.3 Hydrogen Bond 6
  1.3.4 Hydrophobic Interaction 7
  1.3.5 Steric Repulsion 8
1.4 Isomerism: Configuration versus Conformation 8
1.5 Trilogy 11
1.6 References 13

CHAPTER 2 *MONOMER CONSTITUENTS OF BIOMACROMOLECULES* 15

2.1 Nucleotides: Constituents of Nucleic Acids 15
2.2 α-Amino Acids: Constituents of Proteins 18
2.3 Monosaccharides: Constituents of Glycans 23
2.4 Addendum 28
2.5 References 30

CHAPTER 3 *PURIFICATION AND CHARACTERIZATION* 31

3.1 Purification: Overview 31
3.2 Purification: Chromatography 34
3.3 Purification: Electrophoresis 40
3.4 Characterization: General 44
  3.4.1 Purity 44
  3.4.2 Molecular Weight 44
  3.4.3 Molecular Dimension 50
3.5 Characterization: Specific 51
  3.5.1 Melting Temperature of DNA 51
  3.5.2 Buoyant Density of Biomacromolecules 52
  3.5.3 Isoelectric pH of Proteins 52
  3.5.4 Removal of Glycosides from Glycoproteins 53
3.6 References 53

CHAPTER 4 *BIOMACROMOLECULAR STRUCTURE: NUCLEIC ACIDS* 55

4.1 Structural Organization 55
  4.1.1 Structural Hierarchy 55
  4.1.2 Representation of Structures of Nucleic Acids 56
4.2 Sequence Analysis of Nucleic Acids 57
  4.2.1 General 57
  4.2.2 Chemical Cleavage Method 59
  4.2.3 Enzymatic Chain Termination/Dideoxy Method 60
  4.2.4 Mass Spectrometric Analysis 61
  4.2.5 Automated DNA Sequencing Technology 62
4.3 Secondary Structure and Structure Polymorphism of DNA 63
  4.3.1 Key Structural Features of Nucleic Acids 63
  4.3.2 DNA Polymorphism 66
  4.3.3 Alternative Structures of DNA 69
4.4 Supercoiling and Tertiary Structure of DNA 77
  4.4.1 DNA Topoisomers 77
  4.4.2 Superhelical Density and Energetics of Supercoiling. 80
4.5 Classification and Structures of RNA 81
  4.5.1 Structures of RNA 81
  4.5.2 Transfer RNA 82
  4.5.3 Ribosomal RNA 83
  4.5.4 Messenger RNA 84
  4.5.5 Other Classes of RNA 85
4.6 RNA Folds and Structure Motifs 86
  4.6.1 RNA Folds 86
  4.6.2 Structure Motifs of RNA 86
4.7 Energetics of Nucleic Acid Structure 89
4.8 Nucleic Acid Application 90
4.9 References 91

CHAPTER 5 *BIOMACROMOLECULAR STRUCTURE: PROTEINS* 94

5.1 Architecture of Protein Molecules 94
  5.1.1 Introduction 94
  5.1.2 Representation of Protein Structures 94

5.2 Primary Structure of Proteins: Chemical and Enzymatic Sequence Analysis 95
  5.2.1 Amino Acid Composition 96
  5.2.2 Peptide Cleavage, Separation and Analysis 97
  5.2.3 Terminal and Sequence Determination 97
  5.2.4 Peptide Ladder Sequencing 101
5.3 Primary Structure of Proteins: Sequence Analysis by Tandem Mass Spectrometry 101
  5.3.1 An Application of Mass Spectrometry (MS) in Protein Chemistry 101
  5.3.2 Application of Tandem Mass Spectrometry (MS–MS) in Protein Sequence Analysis 103
5.4 Conformational Map 108
5.5 Secondary Structures and Motifs of Proteins 110
  5.5.1 α-Helical Structure 111
  5.5.2 β-Sheet Structure 113
  5.5.3 Nonrepetitive Structure: Connection (Loop) and Turn 115
  5.5.4 Notes to Secondary Structures of Globular Proteins 116
  5.5.5 Motifs: Supersecondary Structures 117
5.6 Domains and Tertiary Structures of Proteins 118
  5.6.1 Domain Structures 119
  5.6.2 Tertiary Structures and Protein Folds 121
  5.6.3 Folds and Protein Binding 126
  5.6.4 Membrane Proteins 128
  5.6.5 Fibrous Proteins 128
  5.6.6 Circular (Cyclic) Proteins 129
  5.6.7 Representation of Protein Topology 130
  5.6.8 Accessible Surface of Folded Structures 130
5.7 Classification of Protein Structures 133
  5.7.1 α-Helical Proteins 133
  5.7.2 β-Sheet Proteins 133
  5.7.3 α + β Proteins 135
  5.7.4 α/β Proteins 135
  5.7.5 Multidomain Structures 135
  5.7.6 Membrane and Cell Surface Proteins 136
  5.7.7 Irregular and Small Proteins 136
5.8 Quaternary (Subunit) Structures of Proteins 137
5.9 Quinternary Structure Exemplified: Nucleoproteins 140
  5.9.1 Chromosomes 140
  5.9.2 Ribosomes 141
  5.9.3 Spliceosome and Splicing Activities 142
5.10 Conformational Energetics 143
5.11 References 144

CHAPTER 6 *BIOMACROMOLECULAR STRUCTURE: POLYSACCHARIDES* 147

6.1 Propagation of Polysaccharide Chains 147
  6.1.1 Introduction 147
  6.1.2 Representation of Glycan Structures 148
  6.1.3 Toward Linear Code for Glycans 148
6.2 Sequence Analysis of Polysaccharides: Primary Structure 153
  6.2.1 Hydrolysis to Constituent Monosaccharides 154
  6.2.2 Chemical Methods 154
  6.2.3 Enzymatic Methods 155
  6.2.4 Spectrometric Methods 157
6.3 Conformation: Secondary and Tertiary Structures of Polysaccharide Chains 161
6.4 Conformation: Description of Some Polysaccharide Structures 163
  6.4.1 Starch 163
  6.4.2 Glycogen 164
  6.4.3 Pectins 165
  6.4.4 Cellulose 165
  6.4.5 Chitin 166
6.5 Glycobiology: Study of Glycoprotein-Associated Glycans 167
  6.5.1 Glycoprotein and Glycoforms 167
  6.5.2 Structure Diversity of Oligosaccharide Chains 168
  6.5.3 Structural Analysis 174
6.6 Neoglycoproteins 177
6.7 Organizational Levels of Biomacromolecular Structures 177
6.8 References 181

CHAPTER 7 *STUDIES OF BIOMACROMOLECULAR STRUCTURES: SPECTROSCOPIC ANALYSIS OF CONFORMATION* 183

7.1 Biochemical Spectroscopy: Overview 183
7.2 Ultraviolet and Visible Absorption Spectroscopy 185
  7.2.1 Basic Principles 185
  7.2.2 Amino Acid Residues and Peptide Bonds 187
  7.2.3 Purines, Pyrimidines and Nucleic Acids 188

- 7.2.4 Perturbation Difference Absorption Spectroscopy 189
- 7.3 Fluorescence Spectroscopy 190
- 7.4 Infrared Spectroscopy 193
  - 7.4.1 Basic Principles 193
  - 7.4.2 Biochemical Applications 195
- 7.5 Nuclear Magnetic Resonance Spectroscopy 197
  - 7.5.1 Basic Principles 197
  - 7.5.2 Two-Dimensional Fourier Transform NMR 202
  - 7.5.3 NMR of Proteins 203
  - 7.5.4 NMR of Nucleic Acids 206
  - 7.5.5 NMR of Glycans 207
- 7.6 Optical Rotatory Dispersion and Circular Dichroism Spectroscopy 208
  - 7.6.1 Basic Principles 208
  - 7.6.2 ORD/CD Spectra and Protein Secondary Structures 209
  - 7.6.3 Empirical Applications of ORD and CD 212
- 7.7 X-ray Diffraction Spectroscopy 214
  - 7.7.1 Basic Principles 214
  - 7.7.2 Crystallographic Study of Biomacromolecules 216
- 7.8 References 219

CHAPTER 8 *STUDIES OF BIOMACROMOLECULAR STRUCTURES: CHEMICAL SYNTHESIS* **220**

- 8.1 Rationale 220
- 8.2 Synthetic Strategy: Conventional Approach 220
  - 8.2.1 Protection and Deprotection of Common Functional Groups 221
  - 8.2.2 Protection and Deprotection Specific to Peptide Synthesis 223
  - 8.2.3 Coupling Reaction 225
- 8.3 Synthetic Strategy: Solid Phase Approach 225
  - 8.3.1 General Concept 225
  - 8.3.2 Solid-Phase Polymer Support 230
- 8.4 Practice of Solid Phase Synthesis and Its Application 232
  - 8.4.1 Oligo- and Polypeptide Synthesis 232
  - 8.4.2 Oligo- and Polynucleotide Synthesis 236
  - 8.4.3 Oligo- and Polysaccharide Synthesis 237
- 8.5 Combinatorial Synthesis 241
  - 8.5.1 Parallel Synthesis 241
  - 8.5.2 Mixture Synthesis 242
- 8.6 Biochemical Polypeptide Chain Ligation 245
- 8.7 References 247

CHAPTER 9 *STUDIES OF BIOMACROMOLECULAR STRUCTURES: COMPUTATION AND MODELING* **249**

- 9.1 Potential Energy and Molecular Thermodynamics 249
- 9.2 Molecular Modeling: Molecular Mechanical Approach 252
  - 9.2.1 Introduction 252
  - 9.2.2 Energy Calculation 254
  - 9.2.3 Energy Minimization 256
  - 9.2.4 Molecular Dynamics 258
  - 9.2.5 Conformational Search 261
  - 9.2.6 Remaining Issues 262
  - 9.2.7 Computational Application of Molecular Modeling Packages 263
- 9.3 Statistical Thermodynamics 264
  - 9.3.1 General Principles 264
  - 9.3.2 Transitions of Regular Structures: Two-State Models 268
  - 9.3.3 Random Structure: Random-Walk Problem 271
- 9.4 Structural Transition: Examples 273
  - 9.4.1 Coil-Helix Transition in Polypeptides 273
  - 9.4.2 Helical Transition in Nucleic Acids 274
  - 9.4.3 Topological Transition of Closed Circular DNA Duplex 276
- 9.5 Structure Prediction from Sequence by Statistical Methods 276
  - 9.5.1 Approaches 276
  - 9.5.2 Secondary Structure of Proteins and Beyond 277
  - 9.5.3 Functional Sites of Proteins 280
  - 9.5.4 Nucleic Acid Fold 281
- 9.6 Molecular Docking: Prediction of Biomacromolecular Binding 282
- 9.7 References 286

CHAPTER 10 *BIOMACROMOLECULAR INTERACTION* **289**

- 10.1 Biomacromolecules in Solution 289
- 10.2 Multiple Equilibria 291
  - 10.2.1 Single-Site Binding 291
  - 10.2.2 Multiple-Site Binding: General 292
  - 10.2.3 Multiple-Site Binding: Equivalent Sites 293

- 10.2.4 Multiple-Site Binding: Nonequivalent Sites **294**
- 10.3 Allosterism and Cooperativity **295**
  - 10.3.1 Models **295**
  - 10.3.2 Diagnostic Tests for Cooperativity **299**
- 10.4 Specificity and Diversity of Antibody-Antigen Interactions **300**
  - 10.4.1 Structure of Antibody **300**
  - 10.4.2 Antibody-Antigen Complex **303**
- 10.5 Complementarity in Nucleic Acid Interactions **305**
  - 10.5.1 DNA-Protein Interaction **305**
  - 10.5.2 Binding of Intercalation Agent to Supercoiled DNA **309**
  - 10.5.3 RNA-Protein Interaction **310**
- 10.6 Molecular Recognition in Carbohydrate-Lectin Interaction **312**
  - 10.6.1 Classification and Structures of Lectins **312**
  - 10.6.2 Lectin-Carbohydrate Recognition: General **315**
  - 10.6.3 Lectin-Carbohydrate Recognition: Ligand Discrimination **318**
- 10.7 References **320**

CHAPTER 11 *BIOMACROMOLECULAR CATALYSIS* **322**

- 11.1 Biocatalyst: Definition and Classification **322**
- 11.2 Characteristics of Enzymes **325**
  - 11.2.1 Enzymes: Catalytic Proteins **325**
  - 11.2.2 Catalytic Efficiency **326**
  - 11.2.3 Enzyme Specificity **328**
  - 11.2.4 Active Site of Enzyme **330**
  - 11.2.5 Multienzyme Complex and Multifunctional Enzymes **331**
- 11.3 Enzyme Kinetics **333**
  - 11.3.1 Fundamental of Enzyme Kinetics **333**
  - 11.3.2 Steady-State Kinetic Treatment of Enzyme Catalysis **336**
  - 11.3.3 Quasi-Equilibrium Treatment of Random Reactions **338**
  - 11.3.4 Cleland’s Approach **339**
  - 11.3.5 Nonlinear Kinetics **339**
  - 11.3.6 Environmental Effects **341**
- 11.4 Enzyme Mechanisms **344**
  - 11.4.1 Essay on Enzyme Reaction Mechanism **344**
  - 11.4.2 Studies of Enzyme Mechanism: Active Site **349**
  - 11.4.3 Studies of Enzyme Mechanism: Transition State **356**
  - 11.4.4 Structure-Activity Relationship **357**
  - 11.4.5 X-ray Crystallographic Studies and Refinement **361**
  - 11.4.6 Case Studies of Enzyme Mechanisms **361**
- 11.5 Enzyme Regulation **374**
  - 11.5.1 Elements of Enzyme Regulation **374**
  - 11.5.2 Covalent Modifications of Enzymes and Cascade Effect **374**
  - 11.5.3 Control of Enzyme Catalytic Activity by Effectors **377**
  - 11.5.4 Structure Basis of Allosteric Regulation: Glycogen Phosphorylase **381**
- 11.6 Abzyme **383**
- 11.7 Ribozyme **386**
  - 11.7.1 Characteristics of Catalytic RNA **386**
  - 11.7.2 Description of Ribozymes **388**
  - 11.7.3 Strategies for Ribozyme Catalysis **392**
- 11.8 References **394**

CHAPTER 12 *SIGNAL TRANSDUCTION AND BIODEGRADATION* **398**

- 12.1 Chemical Transduction: Metabolism **398**
- 12.2 Elements of Signal Transduction **400**
  - 12.2.1 First Messengers **400**
  - 12.2.2 Receptors **400**
  - 12.2.3 Second Messengers **403**
  - 12.2.4 Transducers: GTP-Binding Proteins **403**
- 12.3 Effector Enzymes and Signal Transduction **406**
  - 12.3.1 Adenylyl Cyclase and Signal Transduction **406**
  - 12.3.2 Phospholipase C and Signal Transduction **408**
- 12.4 Topics on Signal Transduction **410**
  - 12.4.1 Calcium Signaling **410**
  - 12.4.2 Phosphorylation and Dephosphorylation in Signaling **414**
  - 12.4.3 Signal Pathways Operated by Receptor Protein Tyrosine Kinase **417**
  - 12.4.4 Signaling Pathways Operated by Nonreceptor Proteins Tyrosine Kinase **419**
- 12.5 Apoptosis **419**
- 12.6 Hydrolysis versus Phosphorolysis of Glycans **422**

12.7 Nucleolysis of Nucleic Acids **424**
12.8 Proteolysis and Protein Degradation **426**
  12.8.1 Proteolytic Mechanism **426**
  12.8.2 Protein Degradation Pathway **427**
12.9 References **433**

CHAPTER 13 *BIOSYNTHESIS AND GENETIC TRANSMISSION* **436**

13.1 Saccharide Biosynthesis and Glycobiology **436**
  13.1.1 Biosynthesis of Biopolymer: Distributive versus Processive **436**
  13.1.2 Biosynthesis of oligo- and polysaccharide chains **436**
  13.1.3 Biosynthesis of Glycoproteins **437**
13.2 Genetic Information and Transmission **442**
13.3 DNA Replication and Repair **445**
  13.3.1 DNA Replication: Overview **445**
  13.3.2 DNA Replication: Enzymology **448**
  13.3.3 Reverse Transcription **455**
  13.3.4 Post-Replicational Modification **456**
  13.3.5 DNA Repair **458**
13.4 Biosynthesis and Transcription of RNA **461**
  13.4.1 RNA Transcription: Prokaryotic System **461**
  13.4.2 RNA Transcription: Eukaryotic System **463**
  13.4.3 Regulation of RNA Transcription **466**
  13.4.4 Posttranscriptional Processing/ Modification **469**
13.5 Translation and Protein Biosynthesis **472**
  13.5.1 Protein Translation: Overview **472**
  13.5.2 Protein Translation: Processes **475**
  13.5.3 Decoding Mechanism **479**
  13.5.4 Recoding, Frameshifting and Expanded Genetic Code **481**
  13.5.5 Rescue System for Stalled Ribosomes **483**
  13.5.6 Posttranslational Modifications of Protein **484**
  13.5.7 Protein Translocation **488**
13.6 Folding of Biomacromolecules **491**
  13.6.1 Overview **491**
  13.6.2 RNA Folding **491**
  13.6.3 *In vitro* Protein Folding Pathway **492**
  13.6.4 Molecular Chaperone in Cytosolic Protein Folding **494**
13.7 Bioengineering of Biomacromolecules **494**
  13.7.1 Recombinant DNA Technology **494**
  13.7.2 RNA Engineering **500**
  13.7.3 Protein Engineering **501**
  13.7.4 Antibody Engineering **506**
13.8 References **511**

CHAPTER 14 *BIOMACROMOLECULAR INFORMATICS* **515**

14.1 Overview **515**
14.2 Biosequences **515**
  14.2.1 Sequencing Biomacromolecules **515**
  14.2.2 Sequence Similarity and Pair-Wise Alignment **517**
  14.2.3 Similarity Search and Multiple Sequence Alignment **522**
  14.2.4 Statistical Significance of Sequence Search/Alignments **524**
14.3 Microarray: General Description **525**
  14.3.1 Introduction **525**
  14.3.2 Surface Preparation for Microarray **525**
  14.3.3 Microarray Targets **528**
  14.3.4 Microarray Probes **529**
  14.3.5 Biochemical Reaction of Microarray **530**
  14.3.6 Microarray Detection **530**
  14.3.7 Data analysis in microarray **531**
14.4 Computer Technology **533**
  14.4.1 Machine: Computer **533**
  14.4.2 Tool: Program, Language and Programming **535**
  14.4.3 Molecular Graphics **537**
  14.4.4 Resource: Internet **540**
  14.3.5 Internet Resources of Biochemical Interest **546**
14.5 Informatics **548**
  14.5.1 Introduction to Database **548**
  14.5.2 Biochemical Databases **549**
  14.5.3 Database Retrieval **551**
14.6 Gene Ontology **553**
14.7 References **555**

CHAPTER 15 *GENOMICS* **558**

15.1 Genome: Features and Organization **558**
  15.1.1 Genome Features **558**
  15.1.2 Gene Mapping **561**
  15.1.3 Information Content of Nucleotide Sequence **563**
  15.1.4 DNA Library **564**
  15.1.5 Alternative Splicing **566**

15.1.6 Gene Variation: Single Nucleotide Polymorphism **567**
15.2 Genome Informatics: Databases and Web Servers **568**
15.2.1 Nucleic Acid Databases **568**
15.2.2 Nucleic Acid Analysis Servers **571**
15.3 Approaches to Gene Identification **571**
15.3.1 Masking Repetitive DNA **575**
15.3.2 Database Searches **576**
15.3.3 Codon Bias Detection **576**
15.3.4 Detecting Functional Sites in the DNA **577**
15.4 Gene Expression **578**
15.4.1 Expression Profiling: DNA Chips **578**
15.4.2 Gene Expression: mRNA Quantification and Transcriptome Analysis **583**
15.5 Genome Project **587**
15.6 References **590**

CHAPTER 16 *PROTEOMICS* **594**

16.1 Proteome: Features and Properties **594**
16.1.1 Proteome Features **594**
16.1.2 Protein Identity Based on Composition and Properties **595**
16.1.3 Physicochemical Properties Based on Sequence **596**
16.2 Proteome Informatics: Sequence Databases and Servers **598**
16.2.1 Amino Acid Sequence **598**
16.2.2 Primary Sequence Database **599**
16.2.3 Secondary Sequence Database **602**
16.2.4 Boutique Databases **605**
16.3 Proteome Informatics: Structure Databases and Servers **605**
16.3.1 Structure Database: Primary Archive **605**
16.3.2 Structure Databases: Substructures and Structure Classification **608**
16.4 Proteome Informatics: Proteomic Servers **610**
16.4.1 Proteome Analysis and Annotation **610**
16.4.2 Integrated Databases **613**
16.4.3 Post-Translational Modifications and Functional Sites **614**
16.5 Protein Structure Analysis Using Bioinformatics **616**
16.5.1 Secondary Structure Predictions **617**
16.5.2 Three-Dimensional Structure Modeling **618**
16.5.3 Sequence Similarity and Alignment **619**
16.5.4 Structure Similarity and Overlap **620**
16.5.5 Fold Recognition and Threading **623**
16.5.6 Homology Modeling **623**
16.5.7 *Ab initio* Prediction of Protein Structure **624**
16.5.8 Solvation **625**
16.5.9 On-line Protein Structure Prediction **626**
16.5.10 Protein–Protein Interaction **628**
16.6 Investigation of Proteome Expression and Function **629**
16.6.1 Two-Dimensional Gel Electrophoresis **629**
16.6.2 Proteome Analysis by Mass Spectrometry **631**
16.6.3 Analysis of Posttranslational Modification by Mass Spectrometry **634**
16.6.4 High Throughput Protein Crystallography **635**
16.6.5 Protein–Protein Interactions by Two-Hybrid Assay **636**
16.6.6 Protein Chip **638**
16.6.7 Activity-Based Probe **640**
16.6.8 Nonsense Suppression Mutagenesis **643**
16.7 Metabolome **647**
16.8 References **650**

CHAPTER 17 *GLYCOMICS* **655**

17.1 Features of Glycomics **655**
17.1.1 Glycobiology: Nomenclature and Representation of Glycans **655**
17.1.2 Glycobiology: Glycoforms **657**
17.1.3 Glycomics: Response to Post-Genomic Era **659**
17.2 Glycomic Databases and Servers **661**
17.2.1 Glycan Structure **661**
17.2.2 Glycan Analysis **663**
17.2.3 Glycosylation of Proteins **665**
17.3 Glycomics: Genetic Approaches **666**
17.4 Glycomics: Proteoglycomic Approaches **668**
17.4.1 Characterization of Glycosylation Sites **668**
17.4.2 Lectin and Glycoenzyme-Based Proteoglycomics **670**

17.4.3 Metabolic Oligosaccharide Engineering **672**
17.4.4 Recombinant Glycoproteins **673**
17.5 Glycomics: Chemoglycomic Approaches **674**
17.5.1 Structural Analysis of Glycans **674**
17.5.2 Glycoprotein Syntheses in Glycomics **674**
17.5.3 Glycochip **675**
17.6 References **678**

CHAPTER 18 *BIOMACROMOLECULAR EVOLUTION* **680**

18.1 Variation in Biomacromolecular Sequences **680**
18.1.1 Mutation as Driving Force of Evolution **680**
18.1.2 Evolutionary Rate and Role of Selection **682**
18.2 Element of Molecular Phylogeny **685**
18.3 Phylogenetic Analysis of Biosequences **687**
18.3.1 General Consideration **687**
18.3.2 Sequence Data **687**
18.3.3 Phylogenetic Method: Distance-Based Approaches **690**
18.3.4 Phylogenetic Method: Character-Based Approaches **691**
18.3.5 Construction of Phylogenetic Tree **692**
18.3.6 Assessment **692**
18.4 Application of Sequence Analyses in Phylogenetic Inference **693**
18.4.1 Phylogenetic Analysis Software **693**
18.4.2 Phylogenetic Analysis with PHYLIP **693**
18.4.3 Phylogenetic Analysis Online **697**
18.5 Evolution of Biosequences **697**
18.5.1 Evolution of Nucleic Acid Sequence **697**
18.5.2 Regulation of Evolutionary Change **699**
18.6 Evolution of Protein Structure and Function **701**
18.6.1 Evolution of Protein Complexity: General **702**
18.6.2 Evolution of Protein Complexity: Domain Duplication **703**
18.6.3 Evolution of Protein Structure: Fold Change **704**
18.6.4 Evolution of Protein Function: Catalytic Site Convergence versus Divergence **706**
18.7 References **708**

INDEX **710**

PREFACE

Biomacromolecules are fundamental structural and functional units of cells and therefore are at the very core of biochemical interest. They have always been the central topics of biochemical texts and literature. Various physicochemical and biochemical investigations greatly improve our knowledge of biomacromolecular structures and dynamics. The computational approach provides a new tool for structural and functional explorations of biomolecules. Sequence analyses and genetic recombination studies have contributed to our understanding of how biomacromolecules function at molecular and genetic levels. Recent years have witnessed an explosion in biological data that are derived primarily from studies of biomacromolecules. An application of information technology to organize, manage, distribute and analyze these biomacromolecular data has ushered in the new discipline of bioinformatics. There is an increased interest and sophistication in the study of biomacromolecules, as genomics and proteomics take central stage of biochemistry, molecular biology and bioinformatics. Slowly though, glycomics has now gained recognition. These developments give rise to a necessity for the comprehensive documentation and unified presentation of the structures, functions and informatics of biomacromolecules, for which this book is proposed to address.

Biomacromolecules, including nucleic acids (polynucleotides), proteins (polypeptides) and glycans (polysaccharides), are either briefly treated in the introductory biochemistry texts or extensively described in the advanced monographs concerning individual classes of compounds. In response to the renewed interest in biomacromolecules among various fields of biomedical sciences, a unified and comprehensive presentation of these topics is needed. The proposed textbook is aimed at bridging the gap between the introductory/elementary biochemistry course and advanced treatises on an individual class of biomacromolecules. The focus is on the integrated presentation of the structural, dynamic and informational biochemistry of nucleic acids, proteins and glycans, not separately, but as combined topics so that their similarities can be identified/acknowledged and differences compared/appreciated. The book intends to meet the demands of students who would like to broaden their biochemical knowledge beyond the introductory level and to prepare those who would like to venture into the advanced field of studies in genomics, proteomics and/or glycomics.

I have been teaching Biomacromolecules ever since the course was introduced into our program in 1968, on-and-off (and mostly on) before and after my retirement. Since the inception, the subject matter has undergone amazing transformations; from generally descriptive to molecular details, from mainly structure/function to informatics. The field has grown to encompass a large volume of information that any attempt to cover even the most superficial aspects of the topics in a single text is practically impossible if not a daunting task.

This book is written for students who have taken elementary/introductory biochemistry and would like to take further courses in biochemistry related to special topics in nucleic acids, proteins, and/or polysaccharides. Thus it is designed for students who are familiar with the general aspects of biochemistry and would like to further their knowledge or for those who contemplate to pursue the field of studies related to biomacromolecules. It serves as an intermediate textbook in biochemistry, molecular biology and bioinformatics. Its content follows the organization of general/introductory biochemistry so that the continuity of biochemical curriculum is preserved; however the focus is on the macromolecular biochemistry. The book is unique in that it treats nucleic acids, proteins and glycans jointly as biomacromolecules and describes their structures, dynamics and informatics together.

Following introductory topics on biomacromolecules (Chapters 1–3), the elements of biomacromolecular structures (Chapters 4–6) and their studies (Chapters 7–9) are presented. The functions of biomacromolecules are discussed in terms of their interactions (Chapter 10), catalyses (Chapter 11) and metabolisms, including genetic transmission and applications (Chapters 12, 13). Biomacromolecular informatics (Chapter 14), namely genomics (Chapter 15), proteomics (Chapter 16) and glycomics (Chapter 16), are introduced. Chapter 18 describes biomacromolecular evolution. Each chapter presents a proper background in structures, dynamics or informatics of biomacromolecules, providing the context for further studies, which is supplemented by a list of references. In the areas where the speed of change and growth is high, a book cannot be either all-inclusive or entirely current. It is especially difficult for an introductory textbook of this nature to cover the topic materials up-to-date and comprehensively. Students are urged to consult the reference materials (literature cited and Web sites) for further understanding.

Balanced approaches include some general descriptions, which have been treated in general biochemistry texts, otherwise serving as introductory to advanced presentations, but however, not dwelling too deeply on the topics of specialized interest. Some materials that are commonly available in general texts are not repeatedly described here so that others may be considered. For example, discussion on the stereochemistry of monosaccharides (Chapter 3), spectral recordings (Chapter 7) and detailed descriptions on physiological functions or transformations of biomacromolecules, are either omitted or briefly mentioned. However, classical approaches of general interest in the study of biomacromolecules are presented, since they may serve as the background knowledge for advancing current understanding. Solid phase synthesis (Chapter 8) used in the fabrication of biochips (Chapter 14), and chemical modification of enzymes (Chapter 11) applied to the design of affinity/activity-based probes (Chapter 16), are some of examples. The choices of materials presented in this text are derived from many years of the author’s teaching experience. The author alone is responsible for inadequate and erroneous presentations that may occur and readers’ suggestions are very much appreciated.

I would like to thank all authors whose published works have contributed to a better understanding of biomacromolecular biochemistry and formed the resource materials of this textbook. The public accessibility of all the sequences and three-dimensional structures of biomacromolecules has greatly facilitated the advancement of our knowledge for the structure, function and informatics of biomacromolecules. The efforts of all the developers, contributors and managers of many outstanding Web sites of biomacromolecules are most appreciated. The writing of this text would not have been possible without the contributions and generosity of these investigators, authors and developers. My wife, Alice, has been most instrumental in helping me complete this text, which I would like to present to her as a gift on our tetracontyl anniversary. It is my pleasure to state that the realization of this text goes to former Editor, Luna Han and Editorial program coordinator, Kristin Hauser. They have patiently urged me to initiate the project prior to their departure for their new posts. I am grateful to the John Wiley staffs, Ian Collins, Thomas Moore, Dean Gonzalez, and Danielle Lacourciere for their timely help to assist the transformation of this manuscript to be publishable and to oversee the completion of this project.

Retired politicians and celebrities write personal memoirs. Retired entrepreneurs and investors write financial guides. Retired engineers and professionals write how-to or do-it-yourself manuals. Why cannot retired academics write text or reference books of their specialized fields? After many years of teaching and research experience, we certainly have lots to write about. I have taken up this project after my mandatory retirement, not without skeptics. However, I am relieved that I have made it.

C. Stan Tsai (stan@tsai-info.com)
Ottawa, Ontario, Canada

ABBREVIATIONS IN REPETITIVE USE

Some abbreviations that appear in the literature, but are not repeatedly used in this text, are mentioned but not listed here.

1D	one-dimensional
2D	two-dimensional
2DE	two-dimensional electrophoresis
2D-PAGE	two-dimensional polyacrylamide gel electrophoresis
3D	three-dimensional
7TM	seven transmembrane
Ψ(I)	hydrophobic (interactions)
A	adenosine/adenylate/adenine
A or Ala	alanine
A or Gal	galactose
AA	Allo A-H
aa-tRNA	3′-*O*-aminoacyl-tRNA, aminoacyl-tRNA
AB	*ab initio* prediction
ABP	activity-based probe
AC	accession number
AC	adenylyl cyclase
ACR	ancient conserved region
AD	transcription activation domain
ADEPT	antibody-directed enzyme prodrug therapy
AD(T)P	adenosine di(tri)phosphate
AFBP	affinity-based probe
AGE	advanced glycated end-product
AIF	apoptosis inducing factor
AISMAG	An Interactive Server-side Molecular Image Generator
ALU	arithmetic logic unit
AN or GalNAc	*N*-acetylgalactosamine
ANS	8-anilino-1-naphthalene sulfonate
AP	apurinic/apyrimidinic
Apaf	apoptotic protease activating factor
APC	anaphase-promoting complex
aRS	aminoacyl-tRNA synthetase
AS	active site/s
ASGPR	asialoglycoprotein receptor
BCM	Baylor College of Medicine
BD	DNA-binding domain
BER	base excision repair
BFGF	basic fibroblast growth factor
bgl	blood group locus (loci)

BHA	benzyhydrylamine
BIND	biomolecular interaction network database
BIOS	basic input/output system
BLAST	Basic Local Alignment Search Tool
BMCD	Biological Macromolecular Crystallization Database
BNL	Brookhaven National Laboratories
Boc	*tert*-butoxycarbonyl
bp	base pair(s)
Bzl	benzyl
C	cytosine
C or Cys	cysteine
CA	correspondence analysis
CAD or CID	collisionally-activated dissociation or collisionally-induced dissociation
cAMP	3′,5′-cyclic adenosine monophosphate
CAPRI	Critical Assessment of Predicted Interactions
CASP	Critical Assessment of Techniques for Protein Structure Prediction
CASPER	Computer Aided Spectrum Evaluation of Regular Polysaccharides
CATH	Class, Architecture, Topology, Homology
CAZYmes	carbohydrate active enzymes
CBP	CREB binding protein
CBS	Center for Biological Sequence Analysis
cc	comment
CCDC	Cambridge Crystallographic Data Centre
cccDNA	covalently closed circular DNA
ccDNA	closed circular DNA
CCSD	Complex Carbohydrate Structure Database
CD	circular dichroism (spectroscopy)
CD	cluster of differentiation
Cdb	cyclin destruction box
CDG	congenital disorders of glycosylation
cDNA	complementary DNA
CDR	complementarity determining region
CDS, cds	coding sequence
CE	cyanoethyl
CERMAV	Centre de Recherches sur les Macromolécules Végétales
CFG	Consortium for Functional Glycomics
CG	cancer gene/s
cGMP	3′,5′-cyclic guanosine monophosphate
CICR	Ca^{2+}-induced Ca^{2+}-release
CID	collision-induced dissociation
cM	centiMorgan
CM	comparative (homology) modeling
CMD	congenital muscular dystrophy
CMR	carbon-13 magnetic resonance (spectroscopy)
CNS	central nervous system
coagF	coagulation factor
Con A	concanavlin A
COSY	J correlated spectroscopy
CP	chlorophenyl

CPG	controlled pore glass
CPK	Corey–Pauling–Koltun
CPS	carbamyl phosphate synthetase
CPU	central processor unit
CRD	carbohydrate-recognition domain
CRE(B)	cAMP-response element (binding protein)
CSA	Catalytic Site Atlas
CSM	common structures of monosaccharides
CSS	Carbohydrate Structure Suite
CT	carbamoyltransferase
CT	consensus trees/supertrees
CU	control unit
Cyt/cyt	cytochrome
D	dihydrouridine
D or Asp	aspartic acid
Da	daltons
(d)A	(deoxy)adenosine
DAG	diacylglycerol
DB	database
DBMS	database management system
(d)C	(deoxy)cytidine/cytidylate
dc	distance calculation
DCA	discriminant correspondent analysis
DCC	dicyclohexyl carbodiimide
DCP	dichlorophenyl
DD	death domain
DDBJ	DNA Data Bank of Japan
DD-PCR	differential display PCR
DE	description/s
DED	death effector domain
(d)G	(deoxy)guanosine/guanylate
DH	dehydrogenase
DIGE	difference in-gel electrophoresis
diS	disaccharides
DKFZ	German Cancer Research Center
DM	distance matrix
DMC	dichloromethane
DMTr	dimethoxy trityl
DNA	deoxyribonucleic acid(s)
DNP	dinitrophenyl
DNS	dansyl
dNTP	deoxyribonucleotide-5′-triphosphate
Dol	dolichol
dosDNA	defined ordered DNA sequences
DPO	dolichol phosphate-oligosaccharide
DP	degree of polymerization
DP	dolichol pyrophosphate
DR	database cross-reference/S
dsDNA	double-stranded DNA
DSS	sodium 2,2-methyl-2-silapentane-5-sulfonate

DT	date of entry
DTT	dithiothreitol
E or Glu	glutamic acid
EA	enzyme-substrate
EBI	European Bioinformatics Institute
EC	Enzyme Commission
EcorL	*Erythrina corallodendron* lectin
EDTA	ethylenediamine tetraacetate
EF	electrofocusing
EFF/FF	empirical force field/force field
EGF	epidermal growth factor
EGFR	epidermal growth factor receptor
EGP	epidermal growth factor
Eif	eukaryotic initiation factor
EMBL	European Molecular Biology Laboratory
EMBnet	European Molecular Biology network
EMin	energy minimization
EP	enzyme-product
EP	eukaryotic primase
EPD	eukaryotic promoter database
EPL	expressed protein ligation
ER	endoplasmic reticulum
ERK	extracellular signal regulated protein kinase
ESI	electrospray ionization
ESP	electrostatic potential
EST	expressed sequence tag
ExPASy	Expert Protein Analysis System
F or Phe	phenylalanine
FAB	fast atom bombardment
FAD	flavin adenine dinucleotide
FaPy	formamidopyrimidine
FB(M)	fragment-based (method)
FF	force field
FG	functional genomics
fibF	fibrinolytic factor
FID	free induction decay
FITC	fluorescein isothiocyanate
FMN/FAD	flavin mononucleotide/flavin adenine dinucleotide
Fmoc	fluorenyl-9-methyloxycarbonyl
FP	fingerprint/s
FP	fluorescent probe
FR	fold recognition
FSSP	fold tree, domain dictionary, sequence neighbors, structure superposition
FT	feature/s
FT	feature table
FT	Fourier transform
FTP	file transfer protocol
G	general
G	guanosine/guanine

G or Glc	glucose
G or Gly	glycine
GA	genetic algorithms
GA	wheat germ agglutinin
Gaba	γ-aminobutyric acid
GAP	GTPase activating proteins
Gase	glycosidase
GAT	glutamine amidotransferase
Gbp	(giga)-base pairs
GBP	glycan binding proteins
GC	gas chromatography
GD(T)P	guanosine di(tri)phosphate
GE	gel electrophoresis
GE	gene expression
GEF	guanine nucleotide exchange
GIF	graphical interchange format
Gla	γ-carboxyglutamate
Glc	D-Glucose
GlcNacT	GlcNAc transferase
GN or GlcNAc	N-acetylglucosamine
GN	gene name
GO	gene ontology
GOA	Gene Ontology Annotation
GOR	Garnier, Osguthorpe and Robson
GP	genome project(s)
gp	grid point
GP	glycogen phosphorylase
GPa	phospho-phosphorylase a
GPb	dephospho-phosphorylase b
GPCR	G-protein-coupled receptor
GPI	glycosylphosphotidylinositol
GR	glutathine reductase
GRE	glucocorticoid response element
GSS	genome survey sequence
GT	glycosylation pathways
GT	glycosyltransferase(s)
GTO	Gaussian type orbitals
GUI	graphical user interface
H or His	histidine
HB	hydrogen bond(s)
hCG	human chorionic gonalotropin
hd	helical domain
HDV	hepatitis delta virus
HEMPAS	hereditary erythroblastic multinuclearity with positive acidified-serum lysis test
HGP	Human Genome Project
HisF	histidine biosynthetic enzyme
HIV	human immunodeficiency virus
HM	histogram matching
HMM	hidden Markow model

hnRNA	heterogeneous nuclear RNA
hnRNP	heterogeneous nuclear ribonucleoprotein
HPLC	high performance (high-pressure) chromatography
HrPAGE	high-resolution polyacrylamide gel electrophoresis
HSE	heat shock element
HSP	heat-shock protein
HSP	high-scoring pair
HTML	HyperText Markup Language
HTPC	high-throughput protein crystallography
HTS	high-throughput screening
hs	heparan sulfate
HTTP	HyperText Transfer Protocol
Hyl (Hk)	hydroxylysine
Hx	hypoxanthine
Hyp (hP)	hydroxyproline
I	invariant method
I or Ile	isoleucine
ICAT	isotope-coded affinity tag
ID	Identifier
IEF	isoelectric focusing
Ig	immunoglobulin
IgG	immunoglobulin G
IGOT	isotope-coded glycosylation-site-specific tagging
IL	interleukin
IMAC	immobilized metal affinity chromatography
IMP	inosine monophosphate
InDel(s)	insertion(s) and/or deletion(s)
INSDC	International Nucleotide Sequence Database Collaboration
insR	insulin receptor
IP	Internet Protocol
IP	inositol phosphate
IPCR	immuno-polymerase chain reaction
IPG	immobilized pH gradient
IPI	International Protein Index
IPL	intein-mediated protein legation
IP_3R	inositol triphosphate receptor
IR	infrared (spectroscopy)
IR	instruction register
IS	insertion sequence
ISOC	Internet society
ISP	Internet service provider
ISREC	Swiss Institute for Experimental Cancer Research
IUBMB	International Union of Biochemistry and Molecular Biology
IUPAC	International Union of Pure and Applied Chemistry
IVS	intervening sequence(s), intron(s)
JPEG	joint photographic experts group
K or Lys	lysine
Kb	thousand base-pairs
Kbp	kilo base-pairs
KcaM	KEGG Carbohydrate Matcher

kDa	Kilo Daltons
KEGG	Kyoto Encyclopedia of Genes and Genomes
KF-Pol I	Klenow fragment of DNA polymerase I
KW	keyword
L or Leu	leucine
LAN	local area network
LC	liquid chromatography
LCD	liquid crystal display
LDH	lactate dehydrogenase
LFA	*Limax flavus* agglutinin
LINE	long interspersed nuclear element
LINUCS	linear notation for unique description of carbohydrate sequence
LOL	*Lathyrus ochrus* lectin
M or Man	mannose
M or Met	methionine
MA	microarray
MAA	*Maackia amurensis* agglutinin
MAG	myelin-associated glycoprotein
MALDI	matrix-assisted laser desorption/ionization
MAP	mitogen activated protein
MAP-KKK	MAP kinase-kinase-kinase
MAR	memory address register
MAS	maskless array synthesizer
Mb	million base-pairs
Mbp	(mega)-base pairs
MBP	mannose binding protein
MBR	memory buffer register
MC	Monte Carlo method
MC-SYM	macromolecular conformations by symbolic programming
MD	mutation data
MEK	MAP (mitogen activated protein) kinase-ERK kinase
MeNPOC	mehylnitropiperonyloxycarbonyl
MHC	major histocompatibility complex
MI	metastable ion
MIME	multipurpose Internet mail extensions
MIP	molecularly imprinted polymer
MIPS	Munich Info Center for Protein Sequences
miRNA	microRNA
ML	maximum likelihood
MLCK	myosin light chain kinase
MM	molecular mechanics
MMDB	molecular modeling database
MMTr	monomethoxy trityl
MO	molecular orbital(s)
MolD	molecular dynamics
momoS	monosaccharides
MPB	mannose binding protein
MPP	mitochondrial processing peptidase
mPu/Py	methylpurin/pyrimidine
MRE	metal response element

mRNA	messenger RNA
MRW	mean residue weight of monomer
mp	matching point
MS	mass spectrometry
ms	mean-square
MS–MS	tandem mass spectrometry
MSP	maximal-scoring segment pair/s
m.u.	mass unit/s
MW	molecular weight
m/z	mass-to-charge ratio
N or Asn	asparagine
N or Neu	neuramic acid
NAC	nascent polypeptide-associated complex
NAD(P)$^+$	nicotinamide adenine dinucleotide (phosphate)
NAD(P)H	reduced nicotinamide adenine dinucleotide (phosphate)
NAPPA	nucleic acid programmable protein array
NAT	natural antisense transcript
NBRF	National Biomedical Research Foundation
NCBI	National Center for Biotechnology Information
NCS	noncrystallographic symmetry
ncRNA	noncoding RNA
NDB	Nucleic Acid Database
NDP	nucleoside diphosphates
NER	nucleotide excision repair
NeuNAc or Sia	N-acetylneuramic acid or sialic acid
NGF(R)	nerve growth factor (receptor)
NIH	National Institute of Health
NJ	neighbor joining method
NK	natural killer
NLM	National Library of Medicine
NMD	nonsense-mediated mRNA decay
NMR	nuclear magnetic resonance (spectroscopy)
NOE	nuclear Overhauser effect
NOESY	nuclear Overhauser effect and exchange spectroscopy
NOS	nitric oxide synthase
NPG	nucleotide phosphoglycose
NR	non-reducing
NR	non-redundant
nrPTK	non-receptor protein tyrosine kinase
nt	nucleotide(s)
Nvoc	6-nitroveratryloxycarbonyl
NW	network
OC	organism classification
oligoS	oligosaccharides
ORD	optical rotatory dispersion (spectroscopy)
ORF	open reading frame
OUT	operational taxonomic unit
OS	operating system
OS	organism species
OST	oligosaccharyltransferase

OUT	operational taxonomic unit/s
P	parsimony
P or Pro	proline
PABA	poly(A) binding protein
PAGE	polyacrylamide gel electrophoresis
PAM	phenylacetamidomethyl
PAM	point accepted mutation
PC	personal computer
PCA	principle component analysis
PCD	programmed cell death
PCNA	proliferating cell nuclear antigen
PCR	polymerase chain reaction
PD	pyrimidine dimer
PDB	Protein Data Bank
PDC	pyruvate dehydrogenase complex
PDGF	platelet derived growth factor
PDP	pyruvate dehydrogenase phosphatase
PEG	polyethylene glycol
PEP	primer extension preamplification
Perl	Practical Extraction and Report Language
PEST	glutamic acid, serine and threonine
PEG	polyethylene glycol
PFGE	pulsed field gel electrophoresis
PEP	primer extension preamplification
PFK	phosphofructokinase
PGGF	platelet derived growth factor
PHP	*Pyrococcus horikoshii*
PHYLIP	Phylogenetic Inference Package
PIC	phenyl isocyanate
PID	phosphotyrosine interaction domain
PIR	Protein Information Resource
PITC	phenyl isothiocyanate
pk	psudoknot
PKA	protein kinase A
PKC	protein kinase C
PL	phospholipase
PLC	phospholipase C
PLD	phospholipase D
PLP	pyridoxal-5′-phosphate
PMR	proton magnetic resonance (spectroscopy)
^{31}PMR	phosphorus magnetic resonance spectroscopy
PMSF	phenylmethylsulfonyl fluoride
PMW	position weight matrix
PNGF	peptide-N-glycosidase F
Pol	DNA polymerase
poly(DA)	poly(deoxyadenylate)
poly(dG-dC)	poly(deoxyguanidylate-deoxycytidylate)
poly(U)	poly(uridylate)
ppm	parts per million
PPP	point-to-point protocol

PPrP	phosphoprotein phosphatase
PrK	protein kinase/s
PSA	Protein Sequence Analysis
PSD	post-source decay
PSD	Protein Sequence Database
pSer/pThr	phosphoserine/phosphothreonine
Ψ	pseudouridine
PSA	prostate-specific antigen
PSI-BLAST	position sensitive iterated-basic linear alignment sequence tool
PT	phenotype/s
PTB	phosphotyrosine binding
PTB	pyrimidine tract-binding proteins
PTC	premature termination codons
PTH	phenylthiohydantoin
PTK	protein tyrosine kinases
PTM	posttranslational modification
PTP	protein tyrosine phosphatase
PTT	protein truncation test
pTyr	phosphotyrosine
PULC	phosphoprotein-ubiquitin ligase complex
PWM	position weight matrix
Q or Gln	glutamine
QM	quantum mechanics
QSAR	quantitative structure-activity relationship
R	purines
R or Arg	arginine
RAAM	reagent array analysis method
RAM	random access memory
RCSB	Research Collaboratory for Structural Bioinformatics
rd	random domain/s
RDRP	RNA polymerase/s
RE	restriction enzyme/endonuclease
RFC	replication factor
R_G	radius of gyration
RISC	RNA-induced silencing complex
RMS or rms	root-mean-square
RMSD	root-mean-square distance/deviation
RNA	ribonucleic acid(s)
RNAi	RNA interference
RNAP	RNA polymerase
Rnase	ribonuclease
RNP	ribonucleoprotein particle
rpHPLC	reverse-phase high performance liquid chromatography
RRM	RNA recognition motif or RNA-binding domain
rRNA	ribosomal RNA
RRS	recombination recognition signal
RT	reverse transcriptase
RT-PCR	reverse transcriptase-polymerase chain reaction
RyR	ryanodine receptor
S or Ser	serine

SA	structural alphabet
SAGE	serial analysis of gene expression
SAM	S-adenosylmethionine
SBA	soybean allutinin
SCF	self-consistent field
scFv	single-chain Fv
SCOP	Structural Classification of Proteins
SDBS	Spectral Database System
SDS	sodium dodecylsulfate
Sec	selenocysteine
SELEX	systematic evolution of ligands by exponential enrichment
SF	solvent flatness
shRNA	short hairpin RNA
SIB	Swiss Institute of Bioinformatics
SINE	short interspersed nuclear element
siRNA	small interfering RNA
SMS	STING Mellennium Suite
SNA	*Sambucus nigra* agglutinin
snmRNA	small nonmessenger RNA
snoRNA	small nucleolar RNA
SNP	single nucleotide polymorphism
snRNA	small nuclear RNA
snRNP	small nuclear ribonucleoprotein
snRNP	small nuclear ribonucleoprotein particles
SOC	store-operated channel
SOM	self-organizing map
SP	Swiss-Prot
SPPS	solid-phase peptide synthesis
SPS	solid-phase synthesis
SRE	serum response element
SRP	signal recognition particle
SRS	Sequence Retrieval System
SSCP	single-strand conformation polymophism
ssDNA	single-stranded DNA
SSB	single strand DNA binding protein
SSE	secondary structure elements
STO	Slater type orbitals
STRP	short tandem repeat polymorphism
STS	sequence tagged site
STV	streptavidin
T	thymidine/thymine
T	transferase(s)
T or Thr	threonine
tBu	*tert*-butyl
TCP	Transmission Control Protocol
TCR	T lymphocyte receptor
TF	transcription factor
TFA	trifluoroacetate(ic acid)
TFMSA	trifluoromethane sulfonic acid
Tg	thymine glycol

Th	Thomson
TH	thiohydratoin
TIFF	tagged image file format
TIM	triosephosphate isomerase
TJA	*Tricosanthes japonicum* agglutinin
TM	tree manipulation
tmRNA	transfer-messenger RNA
TMS	tetramethylsilane
TNR	tumor necrosis receptor
TOF	time-of-flight
TOPS	protein topology cartoons
TP	terminal ptotein
TPP	thiamine pyrophosphate
Tr	triphenylmethyl/trityl
tRNA	transfer RNA/soluble RNA
tRNA/sRNA	transfer RNA/soluble RNA
Ts	tosyl
TSE	transition state ensemble
U	uridine/uridylate
UAS	upstream activating sequence
Ub	ubiquitin
Ubcs	Ub-conjugating enzymes
UPGMA	unweighed pair group method using arithmetic means
URL	Uniform Resource Locator
USB	universal serial bus
UTR	untranslated region
UV	ultraviolet (spectroscopy)
V or Val	valine
VAST	vector alignment search tool
VEGF	vascular endothelial growth factor
VNTR	variable number tandem repeat
W3C	World Wide Web Consortium
W or Trp	tryptophane
WC pair/pairing	Watson–Crick (canonical) pair/pairing
WGS	whole genome shotgun
WPDBL	WPDB loader
WWW	World Wide Web
Y	pyrimidines
Y or Tyr	tyrosine
YPD	Yeast Proteome Database

CHAPTER 1

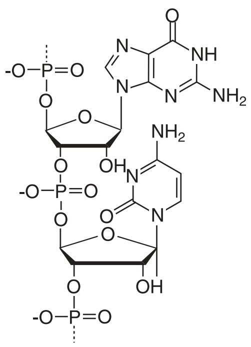
INTRODUCTION

1.1 PRELUDE

A polymer is a large molecule comprised of many fundamental units (known as monomers) joined together. If these units are identical, the result is a homopolymer. If two or more monomer types are involved, the product is a heteropolymer that can be either random or sequential. In the sequential heteropolymers, the monomers are present in the primary structure in a specific sequence. The number of monomer units in a polymer is referred to as the degree of polymerization (DP). When the number is small (2–25), the product is known as an oligomer. Polymers generally refer to molecules with DP >25. In the cells, small building blocks or biomolecular monomers are joined together into biopolymers, also known as biomacromolecules (Jurmark and McFherson, 1984; Korte and Goto, 1976; Walton and Blackwell, 1973). Thus, biomacromolecules (Table 1.1) refer to large molecules of biological interest, with the molecular weights ranging from 10^3 to 10^{12} daltons (Da).

Biomacromolecules are considered as single molecules when they are present in a well-defined stoichiometry and when they display little tendency to dissociate spontaneously under physiological conditions. This text deals specifically with covalent biomacromolecules in which monomer units are linked together by covalent bonds to form giant biomolecules. They include nucleic acids, proteins and polysaccharides (under this definition, biomembranes are excluded). The Web site of the International Union of Biochemistry and Molecular Biology (IUBMB) at http://www.chem.qmw.ac.uk/iubmb/ provides useful information for the nomenclature and conformations of polynucleotides, polypeptides and polysaccharides (Figure 1.1).

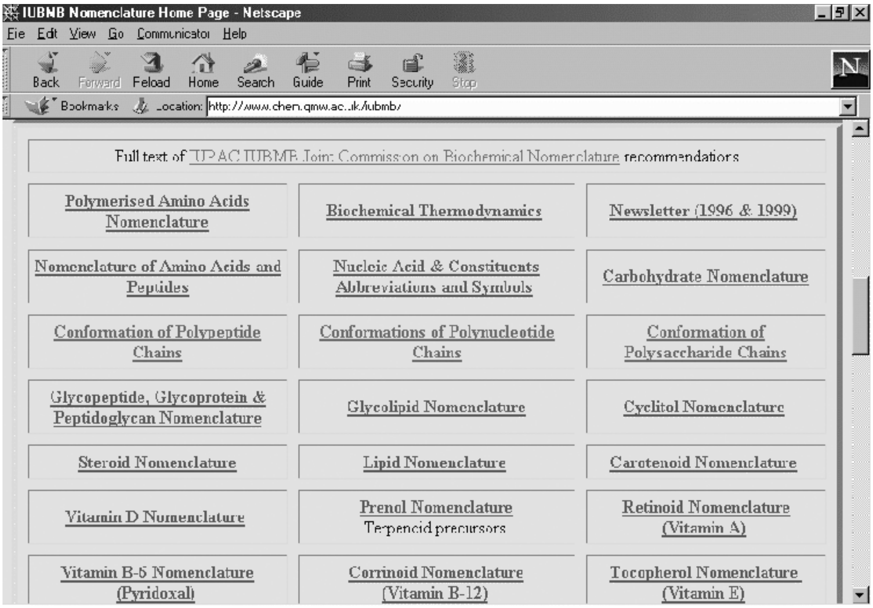
Polynucleotides (Adams *et al.*, 1992; Bloomfield *et al.*, 2000) are long polymers, made up of linear arrays of monomers called nucleotides, consisting of nitrogen bases (pyrimidines and purines) linked to sugar phosphate. The nucleotide units are joined via phosphodiesteric bonds linking the 5′ and 3′ positions of successive sugar residues.



Biomacromolecules, by C. Stan Tsai
Copyright © 2007 John Wiley & Sons, Inc.

TABLE 1.1 Classes of biomacromolecules

Class	Typical molecular weight (Da)	Typical size (nm)	Number of monomers/subunits	Type of monomer/subunit
Oligomers:	10^2–10^4	2	Less than 10^2	
Oligonucleotides				Nucleotides
Oligopeptides				α-Amino acids
Oligosaccharides				monosaccharides
Polymers:	10^4–10^7	4–7	10^2–10^4	
Nucleic acids				Nucleotides
Proteins				α-Amino acids
Polysaccharides				monosaccharides
Supramolecular assemblies:	10^5–10^{12}	20–100	10–10^2	
Nucleoproteins				Nucleic acids, proteins
Protein complexes				Proteins
Glycoproteins				Proteins, glycans
Biomembranes				Lipids, proteins, glycans


IUBMB Nomenclature Home Page - Netscape
File Edit View Go Communicator Help
Back
Forward
Reload
Home
Search
Guide
Print
Security
Stop
Bookmarks
Location: http://www.chem.qmw.ac.uk/iubmb/

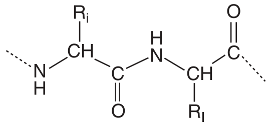
Full text of IUPAC IUBMB Joint Commission on Biochemical Nomenclature recommendations

Polymerised Amino Acids Nomenclature	Biochemical Thermodynamics	Newsletter (1996 & 1999)
Nomenclature of Amino Acids and Peptides	Nucleic Acid & Constituents Abbreviations and Symbols	Carbohydrate Nomenclature
Conformation of Polypeptide Chains	Conformations of Polynucleotide Chains	Conformation of Polysaccharide Chains
Glycopeptide, Glycoprotein & Peptidoglycan Nomenclature	Glycolipid Nomenclature	Cyclitol Nomenclature
Steroid Nomenclature	Lipid Nomenclature	Carotenoid Nomenclature
Vitamin D Nomenclature	Prenol Nomenclature Terpenoid precursors	Retinoid Nomenclature (Vitamin A)
Vitamin B-6 Nomenclature (Pyridoxal)	Corrinoid Nomenclature (Vitamin B-12)	Tocopherol Nomenclature (Vitamin E)

Figure 1.1 Home page displaying a list of information available at IUBMB

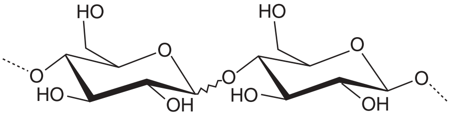
The two types of natural polynucleotides (nucleic acids) are classified according to the sugars they contain. Ribonucleic acid (RNA) contains exclusively β-D-ribose, while the sugar in deoxyribonucleic acid (DNA) is β-2-deoxy-D-ribose. Different nucleic acids can have from around 80 nucleotides (nt), as in transfer RNA (tRNA), to over 10^8 nucleotide-pairs in a single eukaryotic chromosome. The unit for size of nucleic acid is the base (for single-stranded species) or the base-pair (bp, for double-stranded species), with the unit Kb (thousand base-pairs) and Mb (million base-pairs). Examples of synthetic homopolynucleotides are poly(uridylate) or poly(deoxyadenylate), in poly(U) or poly(dA) respectively. A heteropolynucleotide with alternating sequence, poly (dG-dC) is called poly(deoxyguanidylate-deoxycytidylate). If the nucleotides deoxyguanylate (dG) and deoxycytidylate (dC) are randomly distributed over the chain, a comma replaces the hyphen, as in poly(dG,dC). The molecular ratio in the complexes is indicated by a dot between the names, as in poly(A)·2poly(U) for the 1:2 complex of adenylate (A) and uridylate (U).

Polypeptides (Creighton, 1993; Doonan, 2002; Whitford, 2005) are linear polymers of α-amino acids, connected by amide bonds (peptide bonds) between the amino group of one monomer unit and the carboxyl group of the following unit, forming a head-to-tail condensation. One end, called the N-terminal (amino-terminal end), has a free α-amino group, whereas the other end, the C-terminal (carboxyl-terminal end), has a free α-carboxyl group.



Different side chains, R_i, are attached to each α carbon. In addition to the covalent peptide bonds joining adjacent amino acids within a polypeptide chain, covalent disulfide bonds can be formed within the same (intramolecular) polypeptide chain or between different (intermolecular) polypeptide chains. Those polypeptides that occur naturally and have a definite three-dimensional (3D) structure under physiological conditions are called proteins, whilst polymers generated by a random polymerization of amino acids are called polyamino acids (i.e. polylysine). Proteins differ only in the number of amino acids linked together and the sequence in which the various amino acids occur. Natural peptides and proteins encoded by DNA usually contain 20 different L-α-amino acids. These polypeptides range in length from ~40 to over 4000 amino acid residues. Since the average mass of an amino acid residue is ~110 Da, their molecular mass ranges from ~4 to over 440 kDa.

Polysaccharides (Whitford, 2005; Dumitriu, 2005) are composed of many monosaccharide (glycose) units joined together by an acetal linkage (glycosidic linkage) that is formed by the reaction of the hemiacetal hydroxyl group of one unit with an alcohol group of another unit (the wave bond indicates unspecified configuration at the anomeric carbon).



The size of polysaccharide (glycan) molecules can range from 100 to 100 000 monosaccharide units, giving a molecular weight in the range 16–16 000 kDa. Polysaccharides (glycans) containing only one kind of polymerized sugar unit (homoglycans) are more abundant than polysaccharides, which contain two or more kinds of sugar units (hetereoglycans), although the latter are more numerous. Many naturally occurring heteroglycans are the AB glycans composed of repeating sequences of disaccharides. The bond can have either the α- or β-configuration and can belong to any one of the alcohol groups at C-2, C-3, C-4 or C-6 of hexopyranoses. The chain can be either linear or branched in its formation.

The native structures of biomacromolecules are maintained by covalent linkages of their monomer units and various noncovalent interactions between them. Therefore, these structural features (Sheeham, 2000; van Holde, Johnson and Ho, 1998) will be briefly described.

1.2 COVALENT BONDS

Organic molecules, including biomacromolecules, are held together with covalent bonds. Such bonds are usually strong, the standard free energy of formation (ΔG_f^0) being in the order of −400 kJ/mol. Their 3D structures are characterized by bond lengths, bond angles and rotations of groups of atoms about the bonds. The distance between the nuclei of bonded atoms is described as the bond length that corresponds to the sum of their covalent radii (Table 1.2). Thus the C—C single bond has a length of 0.154 nm. The C—O bond is 0.143 nm, the C—H is 0.107 nm and the C—N is 0.147 nm, while the length of a double bond between the two atoms is almost exactly 0.020 nm less than that for a single bond between the same atoms. If there is resonance, that is only partial double bond character, the shortening will be somewhere in between. For example, the C—O distance in the carboxylate anion is 0.126 nm. The distance from the center of an atom to the point at which it contacts an adjacent atom in a packed structure is known as the van der Waals radius. The ways in which biological molecules fit together are determined by these van der Waals contact radii. In every case, the van der Waals radius is approximately equal to the covalent radius, plus 0.080 nm. Van der Waals radii are not as constant as covalent radii because atoms can be squeezed a little, but only enough to decrease the contact radii by 0.005–0.010 nm.

For biomacromolecules, there is usually a tetrahedral arrangement of bonds around single-bonded carbon atoms and phosphorus atoms. All of the bond angles about this carbon atom have nearly the same tetrahedral angle of 109.5°. Bond angles connecting chains of carbon atoms in these compounds vary only slightly from this angle. The contribution of resonance renders the planarity to the double bond with bond angles of ~120°. For example, the peptide linkage is nearly planar, with all the angles falling within 120 ± 4°.

The rotations about a central bond B—C are described by torsion angles involving four atoms in sequence, A—B—C—D. The torsion angle θ is defined as the angle between

TABLE 1.2 The size of common atoms constituting biomolecules

Element	Covalent radius (nm)	van der Waals radius (nm)
H	0.030	0.120
C	0.077	0.160
N	0.070	0.150
O	0.066	0.145
P	0.110	0.190
S	0.104	0.185
	Radius of methyl group	0.200
	Half-thickness of aromatic molecules	0.170

Note: Covalent radii for two atoms can be summed to give the interatomic distance. The van der Waals radii determine how closely molecules can pack. They are usually measured from the shortest distances that can exist between neighboring (and not covalently bonded) atoms in the crystalline state. Thus the closest observed contacts between atoms in macromolecules are approximately 0.02 nm less than the sum of the van der Waals radii.

projected bonds A—B and C—D when looking along the central bond in either direction. It is defined as 0° if A—B and C—D are eclipsed (*cis* and coplanar), and the sign is positive if the far bond is rotated clockwise with respect to the near bond. The torsion angle is usually either in the range 0° to 360° (or −180° to +180°). Another definition uses the dihedral angles, ϕ and φ, that are 'normals' to the two planes formed by A, B, C and B, C, D. In molecules with rotation freedom about single bonds, not all feasible torsion angles are assumed but rather certain sterically allowed conformations are preferred. Therefore it is convenient to describe a structure with torsion angle ranges. The ranges commonly used in organic chemistry (Klyne and Prelog, 1960) are *syn* (~0°), *anti* (~180°), ±*synclinal* (~±60°) and ±*anticlinal* (~±120°). The notation, *cis* (~0°), *trans* (~180°), ±*gauche* (~±60°), is frequently used in crystallographic and spectroscopic literature.

1.3 NONCOVALENT INTERACTIONS

Each newly synthesized biopolymer chain in an undefined structure folds spontaneously to assume its unique native structure. The folding of biopolymer chains, the association of folded biopolymers to form multimeric native macromolecules or assemblies, and the binding of ligands to biomacromolecules proceed without the formation of covalent bonds and are controlled by the noncovalent forces. These noncovalent forces are weak molecular interactions, with energies of formation that are at least one order of magnitude less than that of a covalent bond. They are usually distance-dependent interactions, with the energies being inversely proportional to the distance or some power of the distance separating the two groups (Table 1.3).

It appears that five noncovalent forces are involved in these biochemical processes:

1. electrostatic interaction;
2. van der Waals interaction;
3. hydrogen bond;
4. hydrophobic effect; and
5. steric repulsion.

None of these five categories of noncovalent forces can be completely separated out from all of the others. Van der Waals interactions continue to play a part in each of the other four phenomena. Hydrogen bonds can be considered as a special case of ionic interactions, and the hydrophobic effect tends to reflect hydrogen bonding in the solvent. Therefore, it is informative to discuss each of these categories separately so as to focus on their unique properties.

TABLE 1.3 Relationship of noncovalent interactions to the distance separating the interaction molecules

Type of interaction	Distance relationship
Charge–charge	$1/r$
Charge–dipole	$1/r^2$
Dipole–dipole	$1/r^3$
Charge-induced dipole	$1/r^4$
Dispersion	$1/r^6$
Repulsion	$1/r^{12}$

Note: The distance between the centers of the two interacting atoms is given by r.

1.3.1 Electrostatic interaction (Warshel and Russell, 1984)

The possibility that a positively charged cation might interact favorably with a negatively charged anion and bring two molecules or two parts of the same polymer together is plausible, because unlike charges attract each other. According to Coulomb's law, the interaction between the two unit charges Z_i and Z_j is given by:

$$V = Z_i Z_j e/(Dr_{ij})$$

The interaction as defined by its potential, V is directly proportional to the product of the two changes ($e = 1.602 \times 10^{-19}$ C). V is also inversely proportional to the dielectric constant, D, of the homogeneous medium (D = ~80 for water) and the distance separating the two charged species $r_{ij.}$ The interaction energy between two charges varies as 1/r, which is a long-range interaction. When a positive ion encounters a negative ion in solution, a complex known as an ion pair is formed between these two ions.

1.3.2 Van der Waals interaction (Isrealachvili, 1973)

All atoms and molecules attract each other, even in the absence of charged groups, as a result of mutual interactions related to induced polarization effects. Such polarization is sometimes indicated by symbols δ^+ and δ^-. Attraction results from the uneven distribution of electrons when the positive end of one dipole (polarized bond) is attracted to the negative end of another dipole. This arises from three types of interactions:

1. between two permanent dipoles;
2. between a permanent and an induced dipole; and
3. between two mutually induced dipoles known as London or dispersion forces.

Van der Waals interactions are often represented by the energy potential as a function of distance that includes both the attractive force and the repulsion at close range. The two most commonly used potential functions are the Buckingham potential:

$$V = -A/r_{ij}^6 + B\exp(-\mu r_{ij})$$

and the Lennard–Jones potential:

$$V = -A/r_{ij}^6 + B/r_{ij}^{12}$$

where V is the potential energy due to nonbonded interaction between two atoms (i and j) separated by a distance r_{ij}, and A, B and μ are constants. These two potential functions have been used in most conformational analyses so far performed, although the various workers have differed in their choice of values for the constants, A, B and μ. The first term gives the van der Waals attractions and the second gives the repulsions. The van der Waals interactions break down at distances greater than ~5.0 nm. The optimal distance for the interaction of two atoms is usually 0.03–0.05 nm greater than the sum of their van der Waals radii. Because of the surrounding electron clouds that screen the nuclei, the van der Waals interaction is weak, short in range and fluctuating.

1.3.3 Hydrogen bond (Pimentel and McLellan, 1971)

A hydrogen bond occurs when two electronegative atoms share the same hydrogen atom. The hydrogen atom is formally bonded covalently to the donor atom, but also interacts favorably with the other acceptor atom. The main component of the hydrogen bond is an electrostatic interaction between the dipole of the covalent bond to the hydrogen atom, in which the hydrogen atom has a partial positive charge, and a partial negative charge on the other electronegative atom. The hydrogen atom is able to interact strongly with one electronegative atom while being covalently attached to another, due to its small size and substantial charge, which results from its tendency to be positively polarized. Hydrogen bonds have a strongly directional character and are strongest when the line connecting all three atoms in the bond is straight. The lengths and strengths of hydrogen bonds depend on the electronegativities of the acceptor and donor; the greater the strength of their electronegativities, the shorter the distance between them and the stronger the hydrogen bond itself. In general the donor is an acid and the acceptor is a base, therefore the hydrogen bond is determined by the difference in pK_a between donor and the conjugate acid of the acceptor; the smaller the difference, the stronger the bond. Charged groups also give shorter and stronger hydrogen bonds. Hydrogen bonds are longer than covalent bonds but are shorter than the contact distances given by van der Waals radii. For most types of hydrogen bonds, the bond length between the donor atom and the acceptor atom lies between 0.25 and 0.30 nm. A typical N—H . . . O═C hydrogen bond has the donor nitrogen and acceptor oxygen atoms 0.3 nm apart. The chemical groups in biomacromolecules that most commonly serve as hydrogen bond donors are the N—H and O—H groups. So the most common acceptors are the O═, —O— and —N═ groups. Perhaps the most significant thing about hydrogen bonding is that it often provides the specificity necessary to bring molecules or segments of polymer chains together in complementary ways. In other words, hydrogen bonds are responsible for aligning atoms and holding them at precise distance and angles to each other within the folded structure and complexes.

1.3.4 Hydrophobic interaction (Privalov and Gill, 1988; Muller, 1990)

In an aqueous solution, nonpolar groups and molecules tend to stick together. The relative absence of interactions between nonpolar groups or molecules and water causes interactions among the nonpolar groups themselves to be more favorable than would be the case in other solvents, thus nonpolar molecules greatly prefer nonpolar environment. This preference of nonpolar groups or molecules for nonaqueous environments is known as the hydrophobic interaction. The phenomenon that results principally from the strong internal cohesion of the hydrogen-bonded water structure, is a major contributor to the stability of proteins, nucleic acids and membranes. The hydrophobic interaction can be taken as a transfer of nonpolar molecules or regions of molecules from water into a close association with one another and in which these hydrophobic region are in an environment similar to an inert solvent. The number of water molecules immobilized around the hydrophobic region is decreased. As a consequence, water molecules are freed from the structural area around the hydrophobic regions resulting in an increase in the entropy. The entropy change (ΔS) is usually positive for the hydrophobic interaction between two alkyl or aryl groups. Since $\Delta G = -RT\ln K = \Delta H - T\Delta S$, the free energy for the hydrophobic interaction (ΔG) is negative, i.e. the hydrophobic interaction is strictly an entropy effect. However, as amphipathic (amphiphilic) groups or molecules have both polar and nonpolar regions, the entropy change for hydrophobic interactions may sometimes be zero or even negative, depending on the surrounding water structure. The enthalpy change (ΔH_f) of hydrophobic association of these groups or molecules may be negative enough to make association favorable, such as for the base stacking in DNA.

1.3.5 Steric repulsion

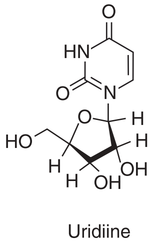
When two atoms or two molecules approach each other, repulsion between them will eventually take place. This steric repulsion, which does not allow two atoms or molecules to occupy the same space at the same time, directly opposes the van der Waals attraction. The repulsive energy is considered to increase with the inverse of the 12th power of the distance between the centers of the two atoms (Lennard–Jones potential). It can also be described as the energy varying exponentially with the inverse of the distance (Buckingham potential). Thus the steric repulsion is a short-range interaction. Because the repulsing energy rises so steeply, it is possible to consider atoms and molecules as having definite dimensions and occupying volumes that are impenetrable to other atoms and molecules. Individual atoms tend to be modeled as spheres and their impenetrable volumes are usually defined by the van der Waals radii. The attractive dispersion and repulsive exclusion together define an optimum distance separating any two neutral atoms at which the energy of interaction is a minimum. Thus this optimum distance defines an effective radius (the van der Waal radius) for each type of atom. Two atoms are generally in close van der Waals contact when the distance between them is approximately 0.08 nm greater than when they are covalently bonded. Optimal van der Waals interactions generally occur at a distance that is about 0.12 nm greater than the covalent length. The van der Waals radius also defines the van der Waals surface area and volume of an atom or molecule (Bondi, 1964).

1.4 ISOMERISM: CONFIGURATION VERSUS CONFORMATION

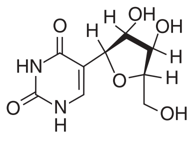
Compounds possessing the same number and kinds of atoms and the same molecular weight (i.e. the same molecular formula) but differing in structure, are called isomers. This phenomenon is called isomerism. Five types of isomerism are recognized:

1. positional isomerism;
2. structural isomerism;
3. geometric isomerism;
4. configurational isomerism; and
5. conformational isomerism.

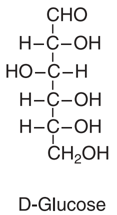
Two compounds can have different structures because of a differing arrangement of the some groups in the positional isomerism (e.g. uridine vs. pseudouridine). Compounds with the same molecular formula but with different functional groups are structural isomers (e.g. D-glucose vs. D-fructose) in structural isomerism. Positional isomers and structural isomers have different chemical and physical properties because of the different arrangement of the atoms. These two types of isomers (i.e. positional isomers and structural isomers), which differ in the manner in which atoms are connected or bonded together, are also called constitutional isomers.



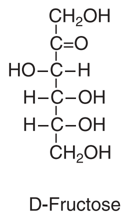
Uridiine



pseudouridine



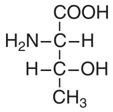
D-Glucose



D-Fructose

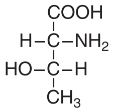
Another type of isomerism is geometric isomerism. While C—C bonds rotate freely, the rotation of C=C bonds requires a larger amount of energy, therefore such a rotation seldom happens at room temperature. This inability of a double bond to rotate is known as hindered rotation, resulting in the formation of the *cis* (Z) isomer and the *trans* (E) isomer. Such pairs of geometric isomers are sometimes called double-bond diastereomers.

The configuration of a molecule refers to the arrangement of its atoms or functional groups in space. Thus the configuration of a molecule defines the position of groups around one or more nonrotating bonds (as double bonds in the case of geometric isomerism) or around a chiral center. Chirality refers to the handedness of an object that is not identical with its mirror image and therefore cannot be superimposed. Such an object must be either asymmetric or dissymmetric. Molecules with the same constitution that differ only in their configurations are called configurational isomers or stereoisomers and lead to stereoisomerism. Stereoisomers are either enantiomers when they are related as a mirror image (e.g. L-threonine and D-threoine) or they are diastereisomers or disateromers (e.g. L-threonine and L-allothreonine). To change the configuration or chirality of a chiral molecule, one bond must be broken to form a planar intermediate, and then the bond must be reformed on the opposite side of the plane. The resulting molecule is the stereoisomer or enantiomer of the starting structure. The mirror-image configurations of the chiral molecule define a pair of enantiomers that have identical chemical and physical properties, except for their opposite but equal optical rotation. Optical isomerism is simply a manifestation of enantiomerism.



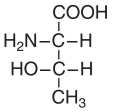
$$\begin{array}{c} \mathrm{COOH} \\ | \\ \mathrm{H_2N-C-H} \\ | \\ \mathrm{H-C-OH} \\ | \\ \mathrm{CH_3} \end{array}$$

L-Threonine
(2S, 3R)-Threonine



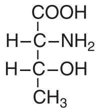
$$\begin{array}{c} \mathrm{COOH} \\ | \\ \mathrm{H-C-NH_2} \\ | \\ \mathrm{HO-C-H} \\ | \\ \mathrm{CH_3} \end{array}$$

D-Threonine
(2R, 3S)-Threonine



$$\begin{array}{c} \mathrm{COOH} \\ | \\ \mathrm{H_2N-C-H} \\ | \\ \mathrm{HO-C-H} \\ | \\ \mathrm{CH_3} \end{array}$$

L-Allothreonine
(2S, 3S)-Threonine

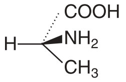


$$\begin{array}{c} \mathrm{COOH} \\ | \\ \mathrm{H-C-NH_2} \\ | \\ \mathrm{H-C-OH} \\ | \\ \mathrm{CH_3} \end{array}$$

D-Allothreonine
(2R, 3R)-Threonine

In biomacromolecules, configuration is most important in describing the stereochemistry of a chiral molecule. A simple chiral molecule (carbon compound) has four unique chemical groups arranged around a tetrahedral carbon atom. Chiral carbon atoms are the most common source of molecular asymmetry and are found in many optically active biological compounds. If there are n asymmetric carbon atoms, then 2^n isomers (stereoisomers) are possible; of these, $2^n/2$ are pairs of enantiomers (mirror images). Stereoisomers, which are not mirror images, belong diastereomers (diastereoisomers). In contrast to enantiomers, the atoms and groups in diastereomers do not have the same relative spatial orientation. Epimers are diasteromers that differ in configuration at only one chiral site of a molecule with multiple chiral centers. For example, L-threonine and L-allothreonine are C-3 epimers. A pair of enantiomers, when present in equal amounts, is called a racemic mixture or racemate. If two of the asymmetric carbons are identical rather than differently substituted within the molecule, a new type of stereoisomer results that is devoid of optical activity and is known as a meso isomer.

The monomer building blocks of biomacromolecules are chiral molecules, with only a few exceptions. The assignment of absolute configurations to biomolecules provides a consistent definition for the configurations of all monomers in a particular class of biomacromolecules. Chiral configurations are designated previously as D and L (using D-glyceraldehyde as a reference) or recently as R and S according to the RS notation (Cahn, 1964). The four groups surrounding the central carbon atom (or other central atom) are ranked according to a priority sequence. The priority of a group is determined by a number of sequence rules, the most important of which is that the higher atomic number precedes the lower. For example, the priorities of the groups in the D-alanine molecule are NH_2 > COOH > CH_3 > H (priority sequence).

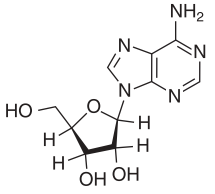


To establish the configuration, we view the molecule down the axis connecting the central atom to the group having the lowest priority (i.e. H). When viewed in this way, the sequence of groups arranged according to the priority sequence can either be that of a right-handed turn (clockwise) as shown in the drawing for D-alanine (R-alanine) or that of a left-handed turn (counterclockwise). The right-handed turn indicates the configuration *R* (rectus); the left-handed turn configuration is *S* (sinister). To establish the priority sequence of groups, we first look at the atoms that are bonded directly to the central atom, arranging them in order of decreasing atomic number. Then if necessary, we move outward to the next set of atoms, again comparing atomic numbers. When double bonds are present at one of the atoms being examined, phantom atoms replicating the real ones are imagined at the ends of the bonds. Some of functional groups encountered in biomolecules are ordered in terms of decreasing priority:

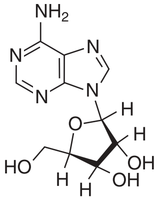
$$SR > SH > OR > OH > NHCOCH_3 > NH_2 > COOH > CHO > CHOH > CH_2OH > C_6H_5 > CH_3 > H.$$

For example, the configuration of L-threonine is 2S, 3R and that of its enatiomer, D-threonine is 2R, 3S. L-Allothreonine with 2S, 3S configuration is its diastereomer. It is noted that a helical chain is chiral, having right-handed (clockwise) and left-handed (counterclockwise) chirality.

Equally important to the stable arrangements of bonded atoms in configuration are conformations. Unlike configurations, the number of possible conformations of a biomacromolecule can be enormous because of the large number of freely rotating bonds. The different spatial orientations that may be assumed by atoms of a molecule as a result of rotation about single bonds are called its conformations. The conformation of a simple molecule can display gauche and anti, eclipsed or staggered conformers (conformational isomers) depending on whether the groups are aligned or misaligned relative to each other on either side of the carbon—carbon bond. For example, the rotation around the glycosyl C1′—N bond renders the nitrogen bases in nucleosides/nucleotides to adopt *anti* in which the bulk of the heterocycles (six-membered ring in purines and O_2 in pyrimidines) point away from the furanose and *syn* in which they turn toward the sugar ring.



anti

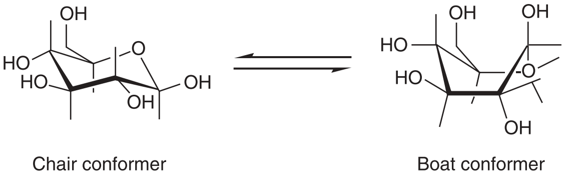


Syn

The preferred conformation will be that in which interactions between atoms on adjacent carbons are kept to a minimum. The conformations of a molecule can be analyzed in terms of three different strain factors:

1. *angle strain*: the deviation from its normal bond angle, for example, 109.5° for tetrahedral carbon;
2. *eclipsing strain*: the rotational preference for a staggered conformation about C—C bonds; and
3. *nonbonded repulsive interactions*: a crowding together of two or more atoms not bonded to each other.

As a rule, molecular species that can interconvert by a pathway having an energy barrier (activation energies) less than 75 kJ/mol cannot be obtained as a discrete substance at room temperature. This is the case for most conformers. However, energy differences of a few kJ/mol suffice to allow a stable chair conformer of pyranoses to dominate over less stable ones at equilibrium in solution (e.g. shown for β-D-glucopyranose).



In contrast, configurational isomers can interconvert only by breaking and reforming covalent bonds.

In essence, constitutional isomers will always have different International Union of Pure and Applied Chemistry (IUPAC) names (http://www.chem.qmw.ac.uk/iupac/), whereas stereoisomers must be identified by an additional nomenclature term. Some overlap exists between configurational isomers and conformational isomers. Configurational isomers are stable under normal conditions whereas interconversion of conformational isomers is often rapid at room temperature. The configurations of biomacromolecules are fixed by covalent bonding. However, the conformations are highly variable and dependent on a number of factors, including the interactions between atoms in the molecule and between the molecule and its environment. The term conformation is generally used to denote secondary and tertiary structures of biomacromolecules.

1.5 TRILOGY

Nucleic acids play a central role in the preservation and transmission of genetic information. DNA is the repository of genetic information that is packaged and organized in higher-ordered forms. Depending on the level of evolution of an organism, the length of the total DNA varies from micrometers to several centimeters. DNA is assembled in simple virus capsids, in prokaryotic cells or in eukaryotic cell nuclei. In human somatic cells there are 46 chromosomes, each consisting of a single DNA duplex molecule about 4 cm long. RNA, which is the generic material of some viruses, is mainly found in the cytoplasm of eukaryotic cells. Proteins are complex macromolecules with exquisite specificity, each being a specialized player in the purposeful and well-controlled activity of the cell. Structurally and functionally, proteins are the most diverse and dynamic of biomolecules, yet linear in their polymeric construction. The relationship reflects the underlying elegant simplicity of the way living systems construct these biomacromolecules, for the nucleic acids that encode the amino acid sequences of proteins are also linear polymers. This permits the direct correspondence between the monomer (nucleotide) sequence of nucleic acid and the monomer (amino acid) sequence of the corresponding polypeptide. By contrast, polysaccharides offer various structural choices, including anomeric configurations, ring structures of monosaccharide units and glycosidic linkages forming linear or branched polymer chains. Table 1.4 provides a snapshot comparison of these biomacromolecules.

The three covalent biomacromolecules share some similarities in their global structures but differ in their cellular functions and structural details. DNA is the carrier of genetic information, which must be faithfully duplicated and selectively transcribed so that each cell can synthesize the proteins it needs. RNA functions mainly in the transcription and translation of this information. In order that the nucleic acids may convey genetic information, a pattern or code must be incorporated in their chemical structure. Purine and pyrimidine bases are arranged in a definite sequence and this sequence constitutes the genetic code.

Proteins are the expression of genetic information and the agents of biological function. An extraordinary diversity of cellular activity is possible only because of the versatility inherent in proteins, each of which is specifically tailored to its biological role. Life forms make use of many chemical reactions to supply themselves continually with chemical energy and to use it efficiently. These reactions in organisms are catalyzed by enzymes, which are proteins. Proteins (e.g. hemoglobin, serum albumine, glucose transporter) transport and control the passaage of biomolecules within the cells or across the membranes. They guide the flow of electrons in the vital process of photosynthesis and electron transfer system coupled to oxidative phosphorylation (e.g. antenna proteins, cytochromes). Protein hormones (e.g. insulin, somatotropin, thyrotropin) transmit information between specific cells and organs in complex organisms. Proteins function as protecting agents (e.g. immunoglobins, thrombin, fibrinogen, antifreeze proteins) of organisms to defend against

TABLE 1.4 Comparison of biomacromolecules

Feature	Nucleic acid	Protein	Glycan
Monomer unit	Nucleotides	Amino acids	Glycoses
Nature of monomer	Compound molecules: consisting of nitrogen base, pentose and phosphate.	Simple molecules: α-amino acids	Simple molecules: glycoses and derivatives (deoxy, N-acetylamino, and carboxyl)
Monomer usage	All 4 major nucleotides	All 20 α-amino acids	Rarely more than 6 glycose types. Mostly one or two.
Residue functionality	Heterocyclic, phosphoesteric	Amino, carboxyl, hydroxy, phenolic thiol, aromatic, heterocyclic	Carbonyl, hydroxy, amino, carboxylic
Stereochemistry of monomer	None for N-bases. Pentoses are β-D-configuration	L (S) configuration	D/L as well as α-/β-anomeric configurations.
Polymer chain	Linear	Linear	Linear and branched
Chain linkage	Phosphoesteric	Amide (peptide)	Acetal (glycosidic)
Chain direction	5′ → 3′	N (amino) → C (carboxyl)	Nonreducing → reducing
Charge at pH 7	Polyanionic	Cationic/anionic/amphoteric	Generally neutral
Biosynthesis	Template (DNA)	Template (mRNA)	Non-template
Biodegradation	Mainly hydrolytic	Hydrolytic	Hydrolytic and phosphorolytic
Major function	Information	Transduction, catalysis, structure	Recognition, structure, food reserve
Informatics	Genomics	Proteomics	Glycomics

intruders or adversaries, and proteins (e.g. lac repressor, catabolite activator protein) control gene expression by binding to specific sequences of nucleic acids. Proteins (e.g. G-actin, myosin, tropomysin) are crucial components of muscles and other systems for converting chemical energy into mechanical energy. They are also necessary for sight, hearing and the other senses. Some proteins (e.g. α-keratin, collagen, elastin) are simply structural, providing the firm architecture within cells and the materials that are used in hair, tendons and bones of animals.

Sequence analysis and structural studies of nucleic acids and proteins have contributed to our understanding of how these biomacromolecules function at the molecular and genetic levels. An application of information technology to collect, analyze, manage and distribute the multitude of data derived from investigations of sequences and structures of nucleic acids and proteins has ushered in the fields of studies in genomics and proteomics collectively known as bioinformatics (Baxevanis and Ouellette, 2005; Tsai, 2002). Links to useful bioinformatics sites are accessible at http://biotech.icmb.utexas.edu/pages/bioinfo.html.

Polysaccharides are components of almost all living organisms. They are most abundant in the higher plants and seaweeds where they constitute approximately three-quarters of the dry weight. Glycans function in two distinct roles; some serve as a means for the storage of chemical energy and others serve as a structural function. Structural glycans are almost always linear molecules, while glycans, which serve primarily as food reserves, are commonly branched or a mixture of linear and branched polysaccharides, with the branched type predominating. In general, branched glycans are easily soluble in water and have immense thickening powers. Linear molecules are excellent structural materials because they pack closely via intermolecular interactions, which make the structure strong, rigid and insoluble. Oligo- and polysaccharides attached to glycoconjugates participate in biological recognition involved in infection, immunity and cell–cell interaction. The synthesis, processing and editing of these glycans are probably mission-oriented. For one purpose, an exactly built glycose is required but for the others, any carbohydrate structure suffices. Glycomics research in the future may provide answers to these important questions (http://www.gak.co.jp/FCCA/).

1.6 REFERENCES

Adams, R.L.P., Knowler, J.T. and Leader, D.P. (eds) (1992) *The Biochemistry of the Nucleic Acids*, 11th edn, Chapman & Hall, London.

Baxevanis, A.D. and Ouellette, B. (eds) (2005) *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, 3rd edn, Wiley-Interscience, New York.

Bloomfield, V.A., Crothers, D.M. and Tinoco, I., Jr. (2000) *Nucleic Acids: Structures, Properties and Functions*, University Science Books, Sansalito, CA.

Bondi, A. (1964) *Journal of Physical Chemistry*, **68**, 441–51.

Cahn, R.S. (1964) *Journal of Chemistry Education*, **41**, 116–25.

Creighton, T.E. (1993) *Proteins: Structures and Molecular Properties*, 2nd edn, W.H. Freeman and Co., New York.

Doonan, S. (2002) *Peptides and Proteins*, Wiley-Interscience, New York.

Dumitriu, S. (ed.) (2005) *Polysaccharide*: *Structure Diversity and Functional Versatility*, 2nd edn, Marcel Dekker, New York.

IUPAC-IUB Commission on Biochemical Nomenclature (1970) *European Journal of Biochemistry*, **17**, 193–201.

IUPAC-IUB Joint Commission on Biochemical Nomenclature (1983) *European Journal of* Biochemistry, **131**, 9–15.

Isrealachvili, J.N. (1973) *Quarterly Review of Biophysics*, **6**, 341–87.

Jurmark, F.A. and McFherson, A. (eds) (1984) *Biological Macromolecules and Assemblies*. John Wiley & Sons, Inc., New York.

Klyne, W. and Prelog, V. (1960) *Experientia*, **16**, 521–23.

Korte, F. and Goto, M. (eds) (1976) *Nucleic Acids, Proteins and Carbohydrates*, Academic Press, New York.

Muller, N. (1990) *Account Chemistry Research*, **23**, 23–8.

Pimentel, G.C. and McLellan, A.L. (1971) *Annual Review of Physical Chemistry*, **22**, 347–85.

Privalov, P.L. and Gill, S.J. (1988) *Pure Applied Chemistry*, **61**: 1097–1104.

Sheeham, D. (2000) *Physical Biochemistry*: *Principles and Applications*, John Wiley & Sons, Inc., New York.

Tsai, C.S. (2002) *An Introduction to Computational Biochemistry*, John Wiley & Sons, Inc., New York.

Van Holde, K.E., Johnson, W.C. and Ho, P.S. (1998) *Principles of Physical Biochemistry*, Prentice Hall, Upper Saddle River, NJ.

Walton, A.G. and Blackwell, J. (1973) *Biopolymer*, Academic Press, New York.

Warshel, A. and Russell, S.T. (1984) *Quarterly Review of Biophysics*, **17**, 283–422.

Whitford, D. (2005) *Proteins: Structure and Function*, John Wiley & Sons, Hoboken, NJ.

World Wide Webs cited

Bioinformatics links: http://biotech.icmb.utexas.edu/pages/bioinfo.html
GlycoForum: http://www.gak.co.jp/FCCA/
IUBMB: http://www.chem.qmw.ac.uk/iubmb/
IUPAC: http://www.chem.qmw.ac.uk/iupac/

CHAPTER 2

MONOMER CONSTITUENTS OF BIOMACROMOLECULES

2.1 NUCLEOTIDES: CONSTITUENTS OF NUCLEIC ACIDS

A variety of organisms collectively synthesize an enormous number of different biomacromolecules (Berg *et al.*, 2002; Voet and Voet, 2004; Garrett and Grisham, 2005), whose great range of physicochemical and biochemical characteristics are derived mainly from the varied properties of their constituent monomer molecules. Therefore some aspects of the monomer chemistry are discussed in this chapter.

The structure and properties of nucleic acids are affected and inherited from the general structure characteristics of their monomer molecules and nucleotides (Blackburn and Gait, 1996; Harmon *et al.*, 1978; Saenger, 1984). Nucleotides are the phosphate esters of nucleosides, which are components of both ribonucleic acid (RNA) and deoxyribonucleic acid (DNA). All nucleotides are constructed from three components:

1. a nitrogen base (purine or pyrimidine);
2. a pentose sugar; and
3. a phosphate residue.

In ribonucleotides, the pentose is the D-ribose while in deoxyribonucleotides, the sugar is 2′-deoxy-D-ribose. The nitrogen base of the planar, heterocyclic molecule derived from purine or pyrimidine is linked to C1′ of the sugar residue. The phosphate group may be bonded to the C3′ or C5′ of a pentose to form its 3′-nucleotide or its 5′-nucleotide respectively. In nucleic acids, the purine and pyrimidine bases of nucleotides serve solely as information symbols for the coding of genetic information.

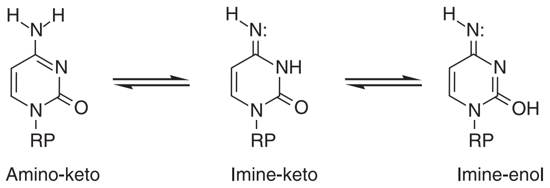
Pyrimidines and purines are characterized by being capable of tautomerism, a phenomenon dependent on ionization whereby the proton becomes attached to one of the interconvertable structures referred to as tautomers. Tautomers can be separated one from the other at low temperature where the rate of interconversion is low. Tautomerization refers to prototropic isomerization that establishes the equilibrium between species, differing in the location of hydrogen and the translocation of a double bond and a single bond. The classic example is the keto-enol equilibrium. Although usually less stable than the keto form, the enol is present in small amounts. It is readily formed from the keto tautomer by hydrogen atoms becoming attached to carbon atoms that are immediately adjacent to carbonyl groups that are remarkably acidic. Easy dissociation of a proton is a prerequisite for tautomerism. Tautomerism is unlikely to occur unless a carbonyl or other activating group is present. Since protons bound to oxygen and nitrogen atoms are usually dissociable, tau-

---

Biomacromolecules, by C. Stan Tsai
Copyright © 2007 John Wiley & Sons, Inc.

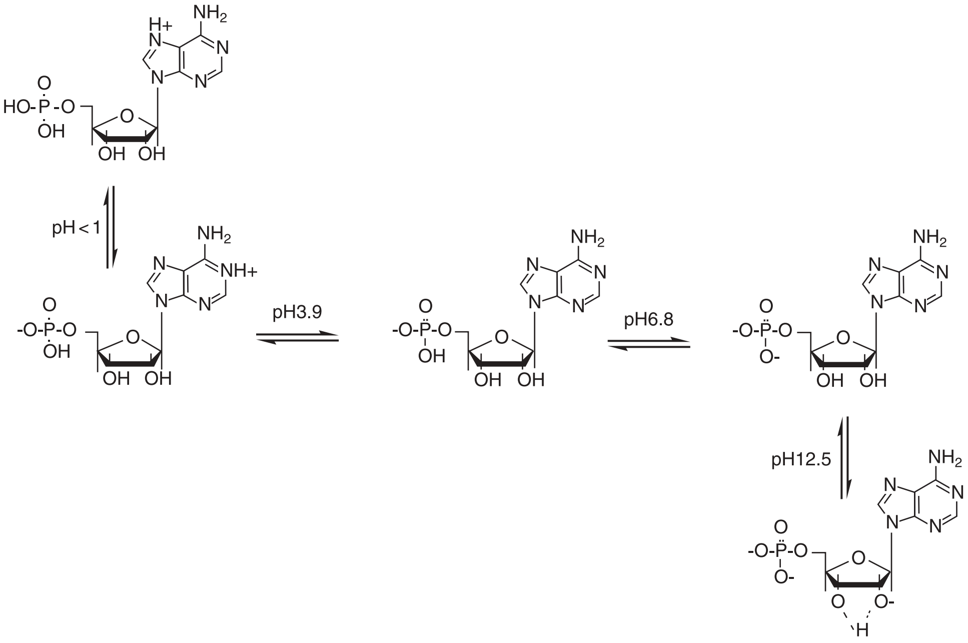
tomerism is possible in amides and in ring systems containing O and N, especially pyrimidine and purine bases.

The choices available to nitrogen bases are the keto-enol and amine-imine equilibrium:



At physiological pH, the five major bases exist overwhelmingly (>99%) in the amino- and keto-tautomeric forms.

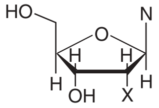
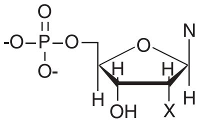
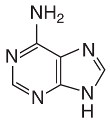
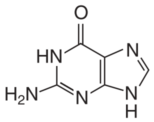
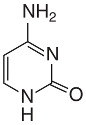
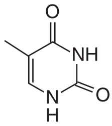
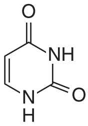
The acid-base behavior of a nucleotide is its most important characteristics. It determines its charge, its tautomeric structure, and thus its ability to donate and accept hydrogen bonds, which is the key feature of base:base recognition. All nitrogen bases in nucleotides are uncharged in the range $5 < pH < 9$. The nucleotide phosphates lose one proton at pH 1 and therefore are monoanionic per nucleotide unit in nucleic acids at the physiological condition.



The pK_a values for the five bases in the major nucleosides and 5′-nucleotides are listed in the Table 2.1. They correspond to the loss of a proton for $pK_a > 9$ and the capture of a proton for $pK_a < 5$. Because the pK_1 value for the first dissociation of a proton from the phosphoric acid moiety is 1.0 or less, the nucleotides have acidic properties.

Another property of pyrimidines and purines is their strong absorbency of ultraviolet light (at ~260 nm), which is also a consequence of the aromaticity of their hetereocyclic ring structures. In addition to the five major bases, their derivatives, known as minor bases,

TABLE 2.1 pK_a values for bases in nucleosides and nucleotides

Base, three-letter and one-letter symbols (site of protonation)	Nucleoside 	5′-Nucleotide 
Purines:		
Adenine, Ade, A (N-1) 	3.52	3.88
Guanine, Gua, G (N-1)	9.42	10.00
Guanine (N-7) 	3.3	3.6
Pyrimidines:		
Cytosine, Cyt, C (N-3) 	4.17	4.56
Thymine, Thy, T (N-3) 	9.93	10.47
Uracil, Ura, U (N-3) 	9.38	10.06

Note: 1. N = Nitrogen base either purine (Pu) or pyrimidine (Py) and X = OH for D-ribose and H for 2-deoxy-D-ribose.

2. For 3′-nucleotides, p is added to the right of one-letter symbols, e.g. Cp for cytidine 3′-monophosphate. For 5′-nucleotides, p is added to the left of the symbols, e.g. pppA for adenosine 5′-triphosphate (ATP).

3. Sequences of nucleotides in nucleic acids are written with one-letter symbols starting with the 5′-terminus at the left toward 3′-terminus at the right. Deoxyribonucleotides in DNA is prefixed with d. If it is not known whether a residue is A or G, the appropriate abbreviation for purine is R; Y is used for pyrimidine, C or T.

4. Special symbols (incompletely specified bases in restriction sequences) include: B for G/T/C (not A), D for A/G/T (not C), H for A,T,C (not G), V for A/G/C (not T) and N for A/C/G/T.

are found in transfer RNAs. The information concerning these minor bases can be accessed at RNA modification database (http://medlib.med.utah.edu/RNAmods).

Nucleotides have compact shapes with several interactions between nonbonded atoms, and their molecular geometry is well reflected in the helical structure of nucleic acids. The shapes of nucleotides can be described in terms of four parameters, namely:

1. *Sugar pucker*: The furanose rings are twisted out of plane in order to minimize nonbonded interactions between their substituents. This puckering is described by identifying the major displacement of C2′ and C3′ of pentose from the median plane of C1′—O4′—C4′. Thus if the *endo*-displacement of C2′ is greater than the *exo*-displacement of C3′, the conformation is called C2′-*endo* and so on. The *endo*-face of the furanose is on the same side as C5′ and the base; the *exo*-face is on the opposite face to the base.
2. *Syn-anti conformation*: The plane of the bases is almost perpendicular to that of the sugar and approximately bisects the O4′—C1′—C2′ angle. This allows the bases to occupy either of two principal orientations. The *anti*-conformer has the smaller H6 (Py) or H8 (Pu) atom above the sugar ring, while the *syn*-conformation has the larger O2 (Py) or N3 (Pu) in that position. Pyrimidines occupy a narrow range of *anti*-conformations, while purines are found in a wider range of *anti*-conformations.
3. *C4′—C5′ orientation*: The orientation of the exocyclic C4′—C5′ bond determines the position of the 5′-phosphate relative to the sugar ring.
4. *C—O and P—O ester bonds*: Phosphate diester is tetrahedral and shows antiperiplanar conformations for the C5′—O5′ bond. X-ray structural studies of tRNA and DNA oligomers suggest that usually H4′—C4′—C5′—O5′—P adopts an extended W-conformation in these structures.

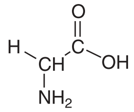
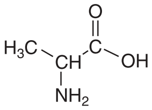
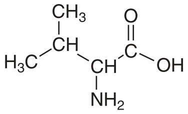
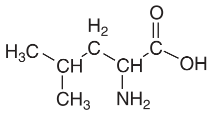
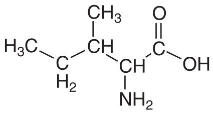
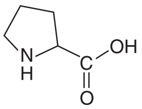
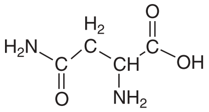
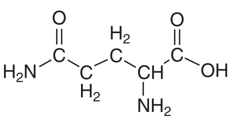
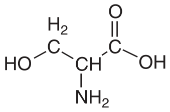
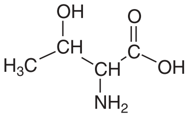
2.2 α-AMINO ACIDS: CONSTITUENTS OF PROTEINS

The 20 α-amino acids vary considerably in their physicochemical properties, such as polarity, acidity, basicity, aromaticity, bulk, conformational flexibility, ability to cross link, ability to hydrogen bond, and chemical reactivity (Barrett and Elmore, 1998; Boulton *et al.*, 1985). These characteristics, many of which are interrelated, are largely responsible for the great range of structures, properties and functions of proteins (Kyte, 1995; Schulz and Schirmer, 1979; Sewald and Jakubke, 2002). α-Amino acids, except for glycine, are chiral molecules and optically active. With four different groups attached, α-carbon is asymmetric, and the configuration about this center for all α-amino acids in proteins is the *L*(*S*-)-stereoisomer, though D-amino acids have occasionally been found in a few naturally occurring peptides. Two α-amino acids, isoleucine and threonine, have a second asymmetric center at the β-carbon; i.e. *S*- for isoleucine and *R*- for threonine. The corresponding amino acids with altered β-carbon stereochemistry are called alloisoleucine and allotheronine.

The α-amino acids in proteins are the most varied of the monomers found in biomacromolecules. These amino acids are distinguished by the chemical properties of their side chains that, in turn, dictate the properties of proteins. Table 2.2 highlights some characteristics of α-amino acids found in proteins.

The physicochemical properties of the side chains of amino acid residues and the amino acid sequence are dominant factors in determining the structure and function of proteins. A number of parameters are proposed to quantify the contribution of each amino

TABLE 2.2 Standard α-Amino acids of proteins and their characteristics

Amino acid (three- and one-letter symbols)	Structural formula	Characteristics
Hydrogen		
Glycine (Gly, G)		With no side chain hindrance, Gly can feed the main chain through tight places in the protein molecule. This gives the polypeptide backbone at Gly residues much greater conformational flexibility.
Alkyl, nonpolar		
Alanine (Ala, A)		The smallest nonpolar residue of Ala does not show preference with respect to the inside or the surface of a protein.
Valine (Val, V)		The nonpolar side chains of Val, Leu and Ile interact favorably with each other and with other nonpolar atoms *via* hydrophobic interaction which is one of the main factors in stabilizing the folded conformations of proteins.
Leucine (Leu, L)		These nonpolar side chains of Leu and Ile are branched with limited internal flexibility. They stiffen the main chain.
Isoleucine (Ile, I)		Nonpolar and branched, the side chain of Ile is involved in hydrophobic interaction which stabilizes the protein conformation.
Proline (Pro, P)		Pro is an imino acid and the imino ring confers conformational stability upon Pro, imposes rigid constraints on rotation about the N-C$^{\alpha}$ bond of the backbone, and disrupts the helical structure. The peptide bond preceding a Pro residue is more likely to adopt the *cis* configuration. Pro is an especially common residue at 'hairpin turns' in globular proteins.
Polar, uncharged		
Asparagine (Asn, N)		These neutral polar residues, Asn, Gln, Cys, Ser and Thr, are found at the surface as well as inside protein molecules. As internal residues they usually form hydrogen bonds with each other or with the polypeptide backbone. Asn is the glycosylation site for oligosaccharides of *N*-glycoconjugates.
Glutamine (Gln, Q)		Asn and Gln are polar and serve both hydrogen-bond donors and acceptors.
Serine (Ser, S)		The primary hydroxyl group of Ser is an excellent nucleophile and it is found at the active sites of many enzymes. Ser serves as the phosphorylation site in various kinase reactions and glycosylation site for *O*-glycoproteins.
Threonine (Thr, T)		The secondary hydroxyl of Thr is polar and nucleophilic but not known to participate in any enzymatic reactions. Thr also provides glycosylation site for *O*-glycoproteins.

(continued)

TABLE 2.2 ***continued***

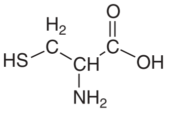
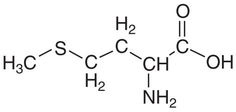
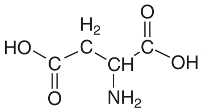
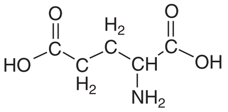
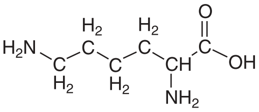
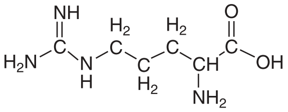
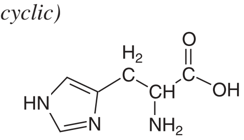
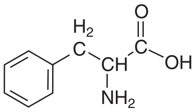
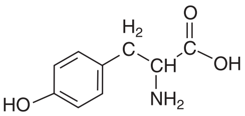
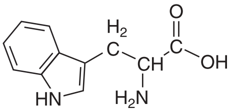
Amino acid (three- and one-letter symbols)	Structural formula	Characteristics
Sulfur containing		
Cysteine (Cys, C)		The thiol group of Cys is the most reactive side residue. The thiolate anion is a potent nucleophile and the thiol is a week acid with $pK_r = 8.37$. Cys serves as the active site residues of many oxidoreductases. Cys residues form complexes of varying stability with a variety of metal ions. It reacts with organic mercurials stoichiometrically. Thiol residues of Cys cross link to form disulfide bonds (cystine) in proteins. Thiols and disulfides undergo rapid exchange and redox reactions.
Methionine (Met, M)		The long side chain of Met is nonpolar. It has a rather flexible side chain containing one sulfur atom in a thioether bond. Met provides the methyl group for many biological methylation reactions *via* formation of S-adenosylmethionine.
Acidic		
Aspartic acid (Asp, D)		The pK_r value for β-COOH of Asp is 3.86. The anions of Asp and Glu play an important role in many enzyme reactions (e.g. glycosidases). The free acid can act as a proton donor and the anion can act as a proton acceptor in acid- and base-catalyzed reactions respectively.
Glutamic acid (Glu, E)		Asp and Glu are negatively charged residues at physiological pH and are found at protein surface. They can be effective chelators of certain metal ions. The pK_r value for γ-COOH of Glu is 4.25.
Basic		
Lysine (Lys, K)		Most of the positively charged Lys and Arg residues are at the molecular surface. PK_r (ε-amino of Lys) = 11.1. The nonionized amino group of Lys is a potent nucleophile. The long arm that bears the ε-amino group serves as a flexible attachment site for certain important cofactors (e.g. FMN, pyridoxal phosphate) of enzymes. ε-Amino of Lys is a target of many nucleophilic modifications.
Arginine (Arg, R)		The Arg side chain consists of the strongly basic δ-guanidino group with $pK_r = 12$. The ionized guanidino group is planar due to resonance. The positively charged Lys and Arg may take part in electrostatic interaction (salt bridge formation). Arg is involved in binding of the phosphate group of nucleotide coenzymes. Arg is susceptible to modification with diketo reagents.
Imidazole (heterocyclic)		
Histidine (His, H)		The imidazole side chain of His resides is a cyclic amine and a potent nucleophile. With a pK_r value of 6.04. The imidazole side chain can be either uncharged or charged and is quite suitable for catalyzing chemical (acid and base) reactions. His is found at the active centers of many enzymes (e.g. hydrolases, transferrases). In the nonionized form, the nitrogen with the hydrogen atom is an electrophile and donor for hydrogen bonding, and the other nitrogen atom is a nucleophile and acceptor for hydrogen bonding. His also participates in chelation of metal ions. His is the target of dye-sensitized photooxidation.

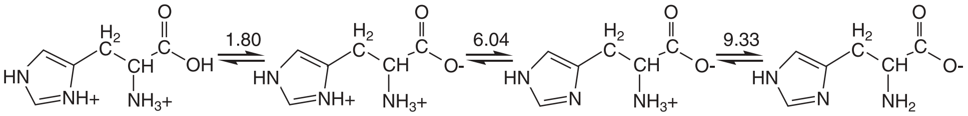
TABLE 2.2 ***continued***

Amino acid (three- and one-letter symbols)	Structural formula	Characteristics
Aromatic, nonpolar		
Phenylalanine (Phe, F)		All aromatic residues, Phe, Tyr and Trp are nonpolar, planar and rather restricted in their side chain flexibility. They are found inside of protein molecules and participate in hydrophobic interactions.
Tyrosine (Tyr, Y)		Tyr possesses a weakly acidic functional group in its aromatic side chain with $pK_r = 10.1$. The phenolic group forms hydrogen bond and the phenolic ring of Tyr is relatively reactive in electrophilic substitution reactions.
Tryptophan (Trp, W)		Trp occurs least frequently in proteins and its indole side chain is the largest side chain. The aromatic side chains are responsible for most of the ultraviolet absorbency and fluorescence properties of proteins. All three possess absorption maxima around 280 nm.

Note: 1. pK_r refers to the ionization constant of the side residues.
2. Sequences of amino acids in proteins are often written with the one-letter symbols, starting with the N-terminal residue, which is at the left and is considered the first residue of the polypeptide chain. Hyphens are inserted between residues to indicate gaps, and other punctuation is used if the sequence is not known entirely. Parentheses around segments of uncertain sequence, dots separating residues whose positions are almost certain, and commas between amino acid residues of unknown sequence. If it is not known whether a residue is Glu or Gln, the appropriate abbreviation is Z, and B is used for Asp or Asn.

acid to the properties and structure of proteins. These parameters can be accessed at AAIndex web site (http://www.genome.ad.jp/dbget-bin/).

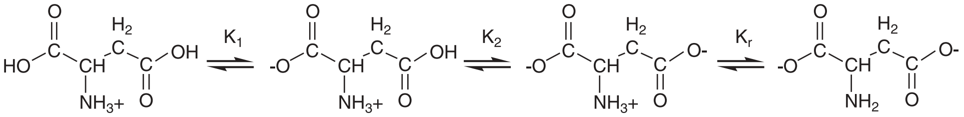
Proteins can be hydrophobic or hydrophilic, though most proteins are amphipathic; that is, they include both hydrophobic and hydrophilic amino acids. Furthermore, proteins are amphoteric by carrying both positive and negative charges. The ionization behavior of α-amino acids is an important property that determines their charge, chemical reactivities and noncovalent interaction patterns. Three types of ionizable groups are recognized in α-amino acids: i) α-carboxylic group; ii); α-amino group; and iii) ionizable side residues.



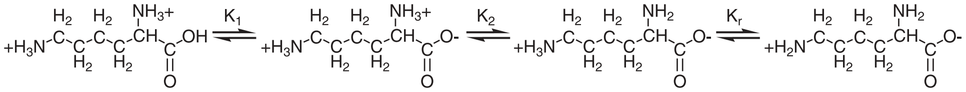
An amino acid can therefore act as either an acid or a base at a certain pH. Compounds with this property are said to be amphoteric and are referred to as ampholytes. Molecules that bear charged groups of opposite polarity are known as zwitterions or dipolar ions. At neutral pH (pH 7), all amino acids carry both positive and negative charges and are therefore zwitterionic. The pH at which the total charge is zero (that is, all the negative charges are balanced by positive charges) is known as the isoelectric point, p*I*, which can be evaluated according to:

$$pI = \tfrac{1}{2}(pK_i + pK_j)$$

where K_i and K_j are the dissociation constants of the two ionizations involving the neutral species. For monoamino-monocarboxylic acids, K_i and K_j represent K_1 and K_2, corresponding to the dissociation constants of α-carboxylic and α-amino groups respectively. Typical values of pK_1 fall between 1.9 and 2.4. Typical values for pK_2 are in the range of 8.7 to 10.1. However, for acidic amino acids, aspartic and glutamic acids, K_i and K_j correspond to K_1 and K_r,



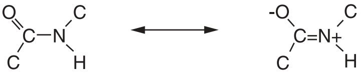
whereas for basic amino acids, arginine and lysine, as well as histidine, these values are K_r and K_2.



K_r is the dissociation constant of the ionizable side residues.

The overall charge of a protein is dependent on the number of acidic and basic amino acids that are charged at a particular pH. Proteins are least soluble in aqueous solution at pH equal to its p*I*. With a large number of diversified α-amino acids as building blocks, nature is able to promote an extremely large variety of combinations, and can thus control the subtle functions performed by proteins.

Peptide bonds are formed between α-amino group and α-carboxyl group of adjacent amino acids. Two resonance structures are feasible for the peptide bond; one in which the π bond links the C and O atoms and the other in which the C and N atoms participate in the π bond:



The resonance stabilization energy is approximately 88 kJ/mol. and the peptide bond is estimated to have 40% double bond character.

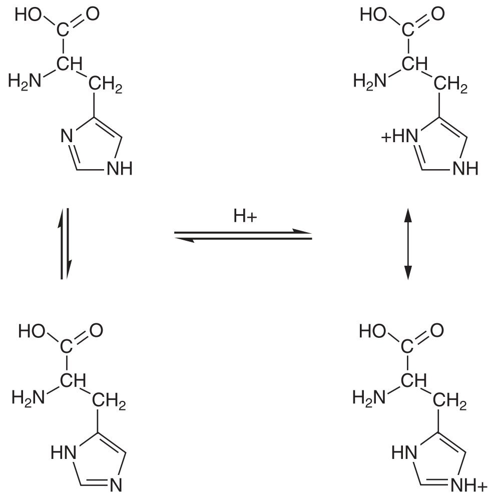
Resonance is a term used to indicate that the properties of a given molecule cannot be represented by a single valence structure but can be represented as a hybrid of two or more structures in which all the nuclei remain in the same places. Only the bonding electrons move to convert one resonance structure into another.

The concept of resonance suggests that if we can draw two or more acceptable structures for a molecular species, then the actual electron does not correspond to any single one of them, but to something between. For this reason these resonance structures are called contributing structures or contributors. The following rules must be observed in using resonance to describe chemical compounds:

- Two or more contributing structures comprise a hybrid structure, which approximates the actual molecule. These contributors are connected by a characteristic double-headed arrow, '←→'.

- The relative positions of atomic nuclei must not change from one contributing structure to another. Only the electrons change location in the contributors that make up the hybrid.
- Resonance is a stabilizing factor. The hybrid of resonance structures (resonance hybrid) is usually more stable than any of its contributing forms by an amount called the resonance energy. The closer the stability of the contributing forms the greater the resonance energy of the hybrid.
- The number of unpaired electrons must remain the same in all contributors.

Resonance and tautomerism are closely related. Thus the acidity of carbon-bound hydrogen in ketones, which allows formation of enol tautomers, is a direct result of the fact that the enolate anion produced by dissociation of one of these hydrogen atoms is stabilized by resonance. Similarly, tautomerism in the imidazole group of histidines is related to resonance in the imidazolium cations.



2.3 MONOSACCHARIDES: CONSTITUENTS OF GLYCANS

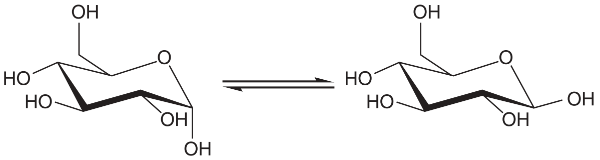
The structures and properties of polysaccharides (glycans) are largely influenced by the types of monosaccharide (glycose) as well as the positions and anomeric configurations that join the polymeric chains (ElKhadem, 1988; Robyt, 1998). Monosaccharides, commonly known as sugars, are aldoses (polyhydroxyaldehydes) and ketoses (polyhydroxyketones) with a molecular formula of $[C(H_2O)]_n$ where usually $n = 3 - 7$. Glycoses contain several chiral centers at secondary hydroxyl carbons, resulting in numerous stereoisomers. The carbonyl functional group forms a cyclic hemiacetal or hemiketal with one of hydroxyl groups (normally C4 or C5) to yield a five-membered furanose ring or six-membered pyranose ring. When hemiacetals and hemiketals are formed, the carbon atom that carries the carbonyl function becomes a chiral carbon atom. Isomers of glycoses that differ only in their configuration about this carbon atom (anomeric carbon) are called anomers, designated as α- or β-anomer. The interconversion between α- and β-anomers via an open chain is accompanied by a change in optical rotation, called mutarotation (Table 2.3).

For example, α-D-glucose (axial hydroxyl at anomeric carbon) has a specific rotation, $[\alpha]_d^{20}$ of 113.2° and that of β-D-glucose (equatorial hydroxyl at anomeric carbon) has

TABLE 2.3 Optical rotations of some naturally occurring hexoses and derivatives

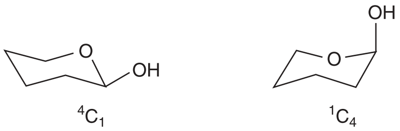
Pyranose	α-Anomer	β-Anomer	Mutaform
D-Glucose	+113	+19	+53
D-Galactose	+150	+53	+80
D-Mannose	+29	−17	+14
D-Galacturonic acid	+107	+31	+52
D-Glucosamine	+100	+25	+73
D-Galactosamine	+121	+45	+95

a specific rotation of 18.7°. The solution of D-glucose reaches an equilibrium value of $[\alpha]_D^{20} = +52.7°$, containing 36.4% of the α anomer and 63.6% of the β anomer.



Not all monosaccharides are found in oligo- and polysaccharides since most glycans contain one or two and rarely more than six types of glycoses. Some glycoses and their derivatives regularly occurring in oligo- and polysaccharides are given in Table 2.4.

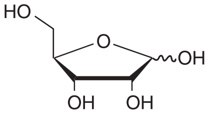
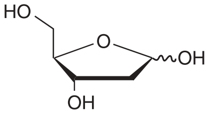
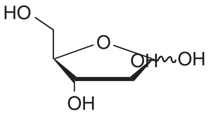
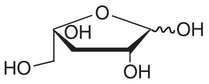
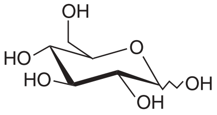
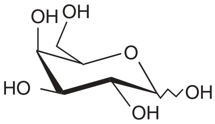
For six-member pyranose rings, six conformations, P (planar), E (envelope), B (boat), S (skew, twisted, twist boat), H (half chair) and C (chair) are possible. The chair conformation C is by far the most stable of the six conformations. Two chair conformations are possible for pyranoses. To distinguish the two conformations, a reference plane is drawn through four atoms (e.g. C2, C3, C5 and O5) of the ring so that the C-atom with the lowest position number (i.e. C1) lies out of the plane. A ring atom that lies above this reference plane is placed before the abbreviation of the ring conformation as a superscript. The one below the plane is written as a subscript following this letter as $^4C_1$ (C4 above and C1 below) and $^1C_4$ (C1 above and C4 below).



Starting with the assumption that all pyranoses exist in a chair form, a hexopyranose will try to adopt a conformation, which allows an equatorial rather than an axial position for the hydroxymethyl (primary hydroxyl) group. Thus in the D-series, a $^4C_1$ conformation will be preferred, except idopyranose in which an equatorial position for the hydroxymethyl group, $^4C_1$, must be at the expense of three axial hydroxyl groups.

Five-member furanose rings are also not planar. The furanose ring can have a common envelope conformation (E), with four atoms coplanar and the fifth out of the plane (any of the four carbons or the oxygen). For the envelop, the conformation is defined by the one atom that is above or below the plane, such as ^3E and E_3 for C3 above and below the plane (C1—C2—C4—O4) respectively. It can also have a less common twist

TABLE 2.4 Common monosaccharides and their derivatives occurring in oligo- and polysaccharides

Monosaccharide (shorthand symbol)	Structural formula	Representative occurrence
Aldopentose		
D-Ribose (D-Rib*f*, B^)		β-D-Rib*f* is the sugar constituent of RNA. It is N-glycosidated with purines (N9) and pyrimidines (N1). Diphosphoesteric chain links C3′ and C5′ of Rib.
2-Deoxy-D-ribose (D-dRib*f*)		β-D-dRib*f* is the sugar constituent of DNA. C1′ of DRib*f* is N-glycosylated with purines (N9) and pyrimidines (N1) while C3′ and C5′ participate in the diphosphoesteric linkage.
D-xylose (D-Xyl*f*, X^)		Xylan in hemicellulose of plants and some marine algae consists of D-Xyl in β-1→4 with branched chain at β-1→3. Rhodymenan (green seaweed) is a linear xylan with β-1→3 linkage. Xyl also occurs in plant gums and mucilages. D-Xyl is attached to threonine in glycoproteins.
L-Arabinose (L-Ara*f*, R)		Arabinan of sugar beets and low-galactan pectins consists of L-Ara*f* in α-1→5 with side chain at α-1→3. Ara coexists with GalA in pectin and Xyl in plant cell walls.
Aldohexose		
D-Glucose (D-Glc*p*, G)		The most abundant glycose, either free or bound, D-Glc*p* is the constituent of many oligo- and polysaccharides including disaccharides, e.g. sucrose, maltose and lactose. Cellulose (plant structural material) and starch (plant food reserve) are two most abundant glycans. Cellulose is a linear homoglucan of D-Glc*p* joined via β-1→4. Starch (plant reserve) is composed of two types of glucans; amylose (Glc in α-1→4) and amylopectin (a mixture of Glc in α-1→4, and ~5% Glc in α-1→6 branch with an average length of 10–13 units). Glycogen (animal reserve) has amylopectin-like structure but higher branched linkage (~10%) with an average length of 20–23 Glc units. Laminaran (brown seaweed, algae, fungi and yeast) has Glc in α-1→3 linkage. Pustulan (lichens) has a linear glucan with D-Glc in β-1→6. Dextran is bacterial glucan that is enzymatically synthesized from sucrose and has α-1→6 linked D-Glc units in the main chain with α-1→2, α-1→3 and α-1→4 branches.
D-Galactose (D-Gal*p*, A)		D-Gal*p* is a constituent of lactose. Galactans of pectin and agar of seaweed consists of Gal in α-1→4 linkage. Galactomannan from endosperm of leguminous seeds is a mixture of Gal and Man in α-1→4 with branch at α-1→6. Gal is also a constituent in oligosaccharides attached to glycoprotein in β-1→4. Agaran (glycan of agar) is composed of alternating sequences of β-1→4 D-Gal and α-1→3 3,6-anhydro-L-Gal. Carrageenan (red seaweed) is a galactan with D-Gal in β-1→3 and α-1→4.

(continued)

TABLE 2.4 ***continued***

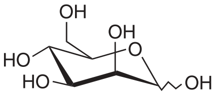
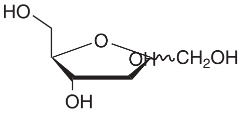
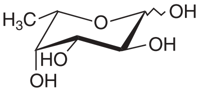
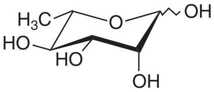
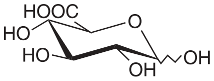
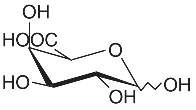
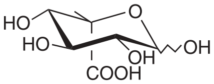
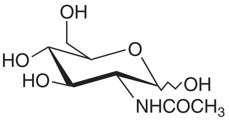
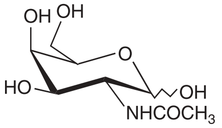
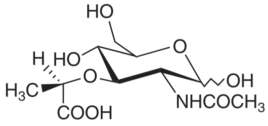
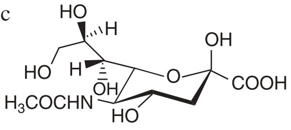
Monosaccharide (shorthand symbol)	Structural formula	Representative occurrence
D-Mannose (D-Man*p*, M)		Mannan from thickened cell walls of palm seeds; yeast is D-Man*p* in β-1→4 and side chain at β-1→6. Glucomannan from tuber consists of Glc and Man. Man is a constituent of oligosaccharides of *N*-glycoproteins in α-1→2, α1→3 and α1→6 (as branching points) and β1→4 (main) linkages. Man is attached to serine in *O*-glycoproteins.
Ketohexose		
D-Fructose (D-Fru*f*, E)		β-D-Fru*f* joins with α-D-Glc*p* via 1↔2 to form sucrose. Fructan (reserve foods in roots, stems, leaves and seeds e.g. inulin, irisan, asparagosan) is D-Fru*f* in α-2→1 and α-2→6 linkages. Levan from various grasses is a linear fructan with D-Fru in β-2→6, and levan from bacteria is branched with D-Fru in β-2→6 and β-2→1.
Deoxyaldohexose		
L-Fucose (L-Fuc*p*, F)		L-Fuc is 6-deoxy-L-Gal. Fucan from brown algae; giant kelp is a polymer of L-Fuc monosulfate. It is a constituent of blood group polysaccharides. As a constituent of glycoconjugates, L-Fuc is linked *via* α-1→4 and α-1→6.
L-Rhamnose (L-Rha*p*, H)		With Man, Gal and GlcNAc, L-Rha form glycan chain of bacterial outer membrane, lipopolysaccharides (*O*-antigen). L-Rha is also found in gums and mucilages.
Monosaccharide derivatives		
D-Glucuronic acid (D-Glc*p*A, U)		Chondroitin and hyaluronic acid are AB glycans composed of repeating disaccharide of D-GlcA and D-GlcNAc with heterolinkages of β-1→3 (GlcA) and β-1→4 (GlcNAc). GlcA is also found in hemicellulose, gums and mucilages.
D-Galacturonic acid (D-Gal*p*A, L)		Pectins are polymers of D-Gal*p*A in α-1→4. PolyGalA that is partially methylated is known as pectinic acids while polyGalA with no or only a negligible amount of methyl ester is called pectic acid. GalA is also a constituent of plant gums and mucilages
L-Guluronic acid (L-Gul*pA*)		Algin (alginic acid) is composed of α-L-Gul*p*A and β-D-Man*p*A (β-D-mannuronic acid) linked by 1→4.
N-Acetyl-D-glucosamine (D-Glc*p*NAc, GN)		Chitin (exoskeleton of insects, crabs, lobsters etc.) is linear homoglycan of D-GlcNac*p* in β-1→4 linkage. GlcNAc is also an alternating constituent of murein in β-1→4. In *N*-glycoproteins, GlcNAc provides attachment to asparagine. GlcNAc is in β-1→2, β-1→4 and β-1→6 linkages. Chondroitin sulfate, heparin, and hyaluronic acid are glycosaminoglycan consisting of repeating disaccharide of D-GlcA and GlcNAc, which is in β-1→4.

TABLE 2.4 ***continued***

Monosaccharide (shorthand symbol)	Structural formula	Representative occurrence
N-Acetyl-D-galactosamine (D-Gal*p*NAc, AN)		D-GalNAc is found in mucopolysaccharide such as chondroitin (cartilage) keratan sulfate (cornea) and dermatan sulfate (skin). In *O*-glycoproteins, GalNAc provides attachment to serine.
N-Acetylmuramic acid (D-Mur*p*Ac)		*N*-Acetylmuramic acid is 2-acetamido-2-deoxy-3-*O*-[(R)-1-carboxyethyl]-D-glucose. MurAc and GlcNAc alternate *via* β-1→4 linkages in murein (peptidoglycan) of bacterial cell wall.
N-Acetylneuramic acid (D-Neu*p*Ac, NN)	 HO H OH HO H O COOH OH H3COCHN HO	N-Acetylneuramic acid, NeuAc is 5-acetamido-3,5-dideoxy-D-glycero-D-galacto-non-2-ulosonic acid. Substituted neuramic acids are also designated as sialic acid (Sia). Sia is linked to oligosaccharide chain of *N*-glycoprotein *via* α-2→3 and α-2→6.

Notes: 1. The shorthand IUPAC nomenclature uses: a) The type of monosaccharide is described by its first three letters except for glucose (Glc), b) The first suffix in *italics* indicates the ring size; *p* for pyranose and *f* for furanose, c) The absolute configurations, D and L, or α and β at the anomeric center is added as a prefix (not shown), d) Derivatives (modifications) are indicated by the last suffix; N for amino, NAc for acetylated amino, and A for carboxyl groups, e) The linkage between two monosaccharides in oligosaccharides or polysaccharides is given with the position numbers and a direction arrow in brackets. If monosaccharides are linked through their anomeric centers, a double-headed arrow is used. For example, lactose as α-D-Gal*p*(1→4)-D-Glc*p*, and sucrose as α-D-Glc*p*-(1↔2)-β-D-Fru*f*, f) Branching of saccharide chain is enclosed within square brackets, [ ], e.g. 1,3-branching of α-D-glucose to maltose as α-D-Glc*p*(1→4)[α-D-Glc*p*(1→3)]-D-Glc*p*.

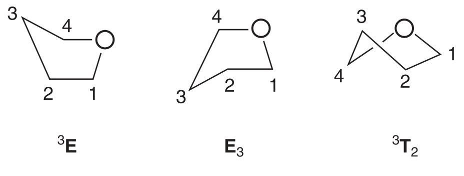
2. For clarity, hydrogens are not explicitly shown, except MurAc and NeuAc where their spatial configurations are shown.

3. The wavy line indicates either α- or β-anomeric bond at the anomeric carbon atom (i.e. unspecified anomeric configurations).

4. For six-membered pyranose ring, all substituents are either equatorial (horizontal bonds radiating from the plane) or axial (vertical bonds perpendicular to the plane).

5. One/two-letter symbols for the proposed linear code (Banin *et al*., 2002) are also included. The basic monosaccharide units are assigned by a single letter code for their common structures (e.g. G for D-Glc*p*, E for D-Fru*f*). Those that are different from the common structure are expressed as follows. a) Opposite stereospecificity to the common structure is indicated with apostrophe “ ’ “, e.g. G′ for L-Glc*p*, R′ for D-Ara*f*, b) Different ring structure to the common structure is indicated with “ ^ “, e.g. G^ for D-Glc*f*, E^ for D-Fru*p*, and c) Differences in both stereospecificity and ring structure is indicate with “ ~ “, e.g. G~ for L-Glc*f*, R~ for D-Ara*p*.

conformation (T) in which three adjacent atoms are coplanar and the other two adjacent atoms are above and below the plane, such as $^3T_2$ in which C3 is above and C2 is below the plane of the other three atoms (C1—O4—C4).



Studies indicate that whereas glucose almost exclusively assumes its pyranose form in aqueous solution, fructose is 67% pyranose and 33% furanose, and ribose is 75% pyranose and 25% furanose, although in polysaccharides, glucose, fructose, and ribose residues are exclusively in their respective pyranose, furanose and furanose forms. The monosaccharide database accessible at http://www.cermav.cnrs.fr/databank/mono/ provides structural information on monosaccharides.

2.4 ADDENDUM

The monomer constituents of biomacromolecules can be separated and identified by various chromatographic methods, which have been employed extensively in biochemical analyses. High performance liquid chromatography (HPLC) is a popular technique for the analysis of small biomolecules (Lim, 1986). An application of HPLC to analyze nucleotides, α-amino acids and monosaccharides is shown in Table 2.5.

Spectroscopic methods are facile, convenient and noninvasive techniques for the identification of chemical compounds. The techniques generally require a small quantity of samples and are particularly suitable for the identification of biomolecules. For this purpose, it is required that the spectra of the biomolecules or related structures of the interested compounds are known. Spectral Database System (SDBS) maintained spectra of many organic compounds, including some biochemical compounds that can be searched and retrieved at http://www.aist.go.jp/RIODB/SDBS/menu-e.html as illustrated in Figure 2.1.

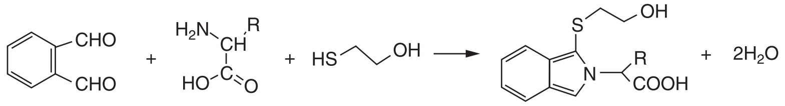
Searches for physical and chemical properties as well as links to useful biochemical and structural sites for these biomolecules can be made at ChemFinder (http://chemfinder.com). The three-dimensional models of biochemical compounds, including nucleotides, α-amino acids, monosaccharides and their derivatives, can be viewed at Klotho (http://www.biocheminfo.org/klotho/). (Figure 2.2)

The monomer biomolecules, though not the main topics of this text, hold a central place in biochemistry. They are intermediates between 'living' and 'non-living' phenomena of chemical compounds and precursors of biomacromolecules, which are the functional units of living systems. Various reactions are involved in the breakdown of biomacromolecules into monomer constituents and their further degradations (catabolism),

TABLE 2.5 Examples of HPLC analyses of nucleotides, α-amino acids and monosaccharides

	Nucleotide	α-Amino acid	Monosaccharide
Sample	Nucleotides	o-phthaldialdehyde (OPT)-amino acids	Glycoses and glycosamines
Column (size), immobile phase	APS-Hypersil (0.46 × 10 cm)	ODS-Hypersil (0.4 × 15 cm)	APS-Hypersil (0.46 × 25 cm)
Elution	Gradient	Gradient	Gradient
Eluent, mobile phase	0.04 M KH_2PO_4, pH2.8 to 0.05 M KH_2PO_4 + 0.8 M KCl, pH2.7	A. 80% 0.05 M Na-PB (pH5.5) in methanol B. 20% 0.05 M Na-PB (pH5.5) in methanol 0–10% B for 10 min 10–85% B for 30 min 85 and 0% B for 5 and 10 min	Acetonitrile : dil.HCl, pH3.0 75 : 25 (v/v) to 65 : 35 (v/v) for 10 min, then 35 : 65 (v/v) for 20 min
Detection	UV at A_{254}	Fluorimetric, electrochemical	Refractive index

Notes: 1. OPT derivation of amino acids follows:



2. APS- and ODS-Hypersils are aminopropylsilyl- and octadecylsilyl-Hypersil (5 μm spherical silica).

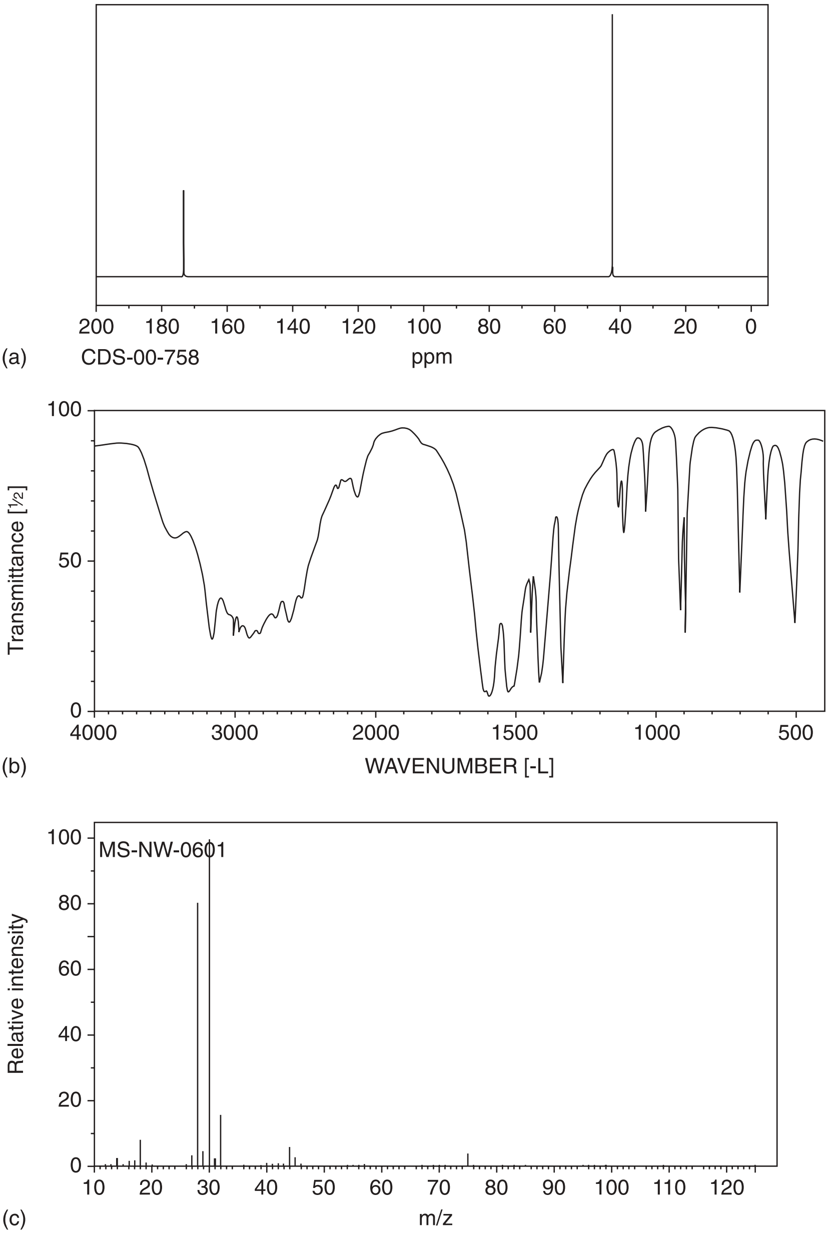


Figure 2.1 Spectra of retrieval from SDBS
The retrieved spectra are C^{13}-nuclear magnetic resonance (a), infrared (b) and mass (c) spectra of glycine.

while elaborately regulated reaction sequences are involved in the assembly of monomer building blocks to form biomacromolecules (anabolism). These biochemical reactions are catalyzed by enzymes and their full descriptions can be obtained from standard biochemistry texts (Berg *et al.*, 2002; Garrett and Grisham, 2005; Hames and Hooper, 2005; Mathews *et al.*, 2000; Voet and Voet, 2004). Kegg Web site at http://www.kegg.com/ provides relevant information about anabolic and catabolic processes known as metabolic pathways and metabolites.

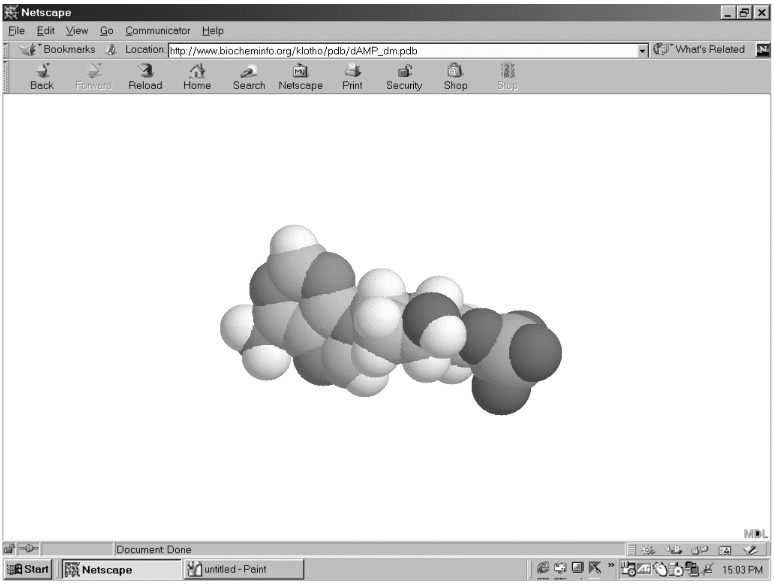

Netscape
File Edit View Go Communicator Help
Bookmarks
Location: http://www.biocheminfo.org/klotho/pdb/dAMP_dm.pdb
What's Related
Back
Forward
Reload
Home
Search
Netscape
Print
Security
Shop
Stop
MDL
Document: Done
Start
Netscape
untitled - Paint
15:03 PM

Figure 2.2 The CPK model of dAMP viewed at Klotho

2.5 REFERENCES

Banin, E., Neuberger, Y., Altshuler, Y. *et al.* (2002) *Trends Glycoscience Glycotechology*, **14**,127–37.

Barrett, G.C. and Elmore, D.T. (1998) *Amino Acids and Peptides*, Cambridge University Press, Cambridge, UK.

Berg, J.M., Tymoczko, J.L. and Stryer, L. (2002) *Biochemistry*, 5th edn, W.H. Freeman, New York.

Blackburn, G.M. and Gait, M.J. (1996) *Nucleic Acids in Chemistry and Biology*, 2nd edn, Oxford University Press, Oxford, UK.

Boulton, A.A., Baker, G.B. and Wood, J.D. (eds) (1985) *Amino Acids*, Humana Press, Clifton, NJ.

ElKhadem, H.S. (1988) *Carbohydrate Chemistry: Monosaccharides and their Oligomers*, Academic Press, San Diego, CA.

Garrett, R.H. and Grisham, C.M. (2005) *Biochemistry*, 3rd edn, Thomson Brooks/Cole, Belmont, CA.

Hames, D. and Hooper, N. (2005) *Biochemistry*, Taylor & Francis, New York.

Harmon, R.E., Robin, R.K. and Townsend, L.B. (eds) (1978) *Chemistry and Biology of Nucleosides and Nucleotides*, Academic Press, New York.

Kyte, J. (1995) *Structure in Protein Chemistry*, Garland Publishing Inc., New York.

Lim, C.K. (ed.) (1986) *Hplc of Small Molecules*: *A practical approach*, IRL Press, Oxford, UK.

Mathews, C.K., van Holde, K.E. and Ahern, K.G. (2000) *Biochemistry*, 3rd edn, Benjamin Cummings, San Francisco, CA.

Robyt, J.F. (1998) *Essentials of Carbohydrate Chemistry*, Springer-Verlag, New York.

Saenger, W. (1984) *Principles of Nucleic Acid Structure*, Springer-Verlag, New York.

Schulz, G.E. and Schirmer, R.H. (1979) *Principles of Protein Structure*, Springer-Verlag, New York.

Sewald, N. and Jakubke, H.-D. (2002) *Peptides*: *Chemistry and Biology*, Wiley-VCH, New York.

Voet, D. and Voet, J.G. (2004) *Biochemistry*, 3rd edn, John Wiley & Sons, Inc., New York.

World Wide Webs cited

AAIndex:	http://www.genome.ad.jp/dbget-bin/
ChemFinder	http://chemfinder.com
KEGG, biosynthetic pathways:	http://www.kegg.com/
Klotho of biochemical compounds:	http://www.biocheminfo.org/klotho/
Monosaccharide database:	http://www.cermav.cnrs.fr/databank/mono/
RNA modification database:	http://medlib.med.utah.edu/RNAmods
Spectral Database System:	http://www.aist.go.jp/RIODB/SDBS/menu-e.html.

CHAPTER 3

PURIFICATION AND CHARACTERIZATION

3.1 PURIFICATION: OVERVIEW

Investigation of biomacromolecules involves their purification, either from biological sources or products derived from chemical synthesis or recombinant technology. It is perhaps the most demanding, laborious and yet essential step preceding structural and functional studies of biomacromolecules (Harris and Angal, 1989). Since laboratory scientists desiring specific biomacromolecules generally design purification processes, this topic has rarely been discussed in introductory texts. However, the purification of biomacromolecules is the most problematic encountered by practicing biochemists. This chapter presents an overview of techniques for the purification and characterization of these biomolecules.

There are usually a few necessary main steps in the isolation/purification procedures of biomacromolecules, including:

- preparation of cell-free extract: cell disruption and separation of subcellular organelles;
- concentration or the preparation and pretreatment to remove major contaminants;
- primary purification, removal of remaining contaminants;
- high-resolution purification;
- polishing of final product.

Protocols for the purification of biomacromolecules (Aboul-Enein, 1999), especially nucleic acids (Bowien and Durre, 2003) and proteins (Deutscher, 1990; Rosenberg, 2005), can be accessed from various monographs and the Web site at http://www.bio.com/protocolstools/protocol.jhtml. Laboratory calculations related to purifications of biomacromolecules are available at LabVelocity (http://reserchlink.labvelocity.com/tools/index.jhtml).

The first step is not required for chemically synthesized products, otherwise prior cell disruption and organelle separation are required to yield cell-free extract from which the desired biomacromolecule can be purified from its natural environment. Total cellular or tissue proteins may be solubilized and assayed prior to purification (Shaw, 1998). Different approaches are available to lyse cells (http://expasy.cbr.nrc.ca/ch2d/protocols/). An approach can be as gentle as adding a surfactant or subjection to an osmotic shock, or can be more energetic such as ultra-sonification, bead beater or French press. Table 3.1 lists some of common methods for cell disruption.

Differential centrifugation or density gradient centrifugation generally accomplishes the subcellular fractionation of organelles. Figure 3.1 shows general guidelines for the

---

Biomacromolecules, by C. Stan Tsai
Copyright © 2007 John Wiley & Sons, Inc.

TABLE 3.1 Methods for cell disruption

Method	Principle/procedure
Cell lysis or autolysis	Osmotic disruption of cell membrane under high osmotic pressure (usually 20% sucrose) or detergent. Autolysis (e.g. yeast) is generally carried out in the presence of toluene.
Freeze and thawing	Repeated cycles of freeze (in liquid nitrogen) and thawing to facilitate disruption of cell membrane.
Enzymatic digestion	Digestion of microbial cell wall with glycanases (e.g. lysozyme for bacteria and β-1,3-glucanase for yeast cells) followed by osmotic release of cell contents.
Homogenization	Soft animal tissues are homogenized in hypotonic buffer using a Potter-Elvenhjem or Dounce homogenizer by forcing cells through narrow gap between pestle and vessel.
Blendering	Animal tissues are homogenized in Waring blender where cells are broken and sheared by rotating blades.
Grinding	Plant materials can be homegenized by grinding with acid washed quartz sand (0.5 g/mL) in a mortar and pestle or in a Waring blender.
Mechanical lysis	Bacterial and yeast cell walls are disrupted by agitation with abrasives using glass beads (0.2 mm diam.) in a mill or by shearing through the small orifice of a French press at very high pressure.
Explosive decompression	Cells equilibrated with nitrogen gas at high pressure and disruption occurs upon their sudden release into atmospheric pressure.
Ultrasonication	High pressure sound waves cause cell breakage by cavitation and shear forces.

Notes: To purify nucleic acids and proteins, an initial subcellular fractionation may be advantageous. In this case the lysis buffer (e.g. 50 mM tris-HCl, pH7.5) generally contains dithiothreitol (DTT), ethylenediamine tetraacetate (EDTA) and/or protease inhibitors such as phenylmethylsulfonyl fluoride (PMSF) in isotonic sucrose (0.25 M sucrose, d = 1.0317). To purify polysaccharides, cells may be lysed/extracted with organic solvents (e.g. n-butanol) or alkali.

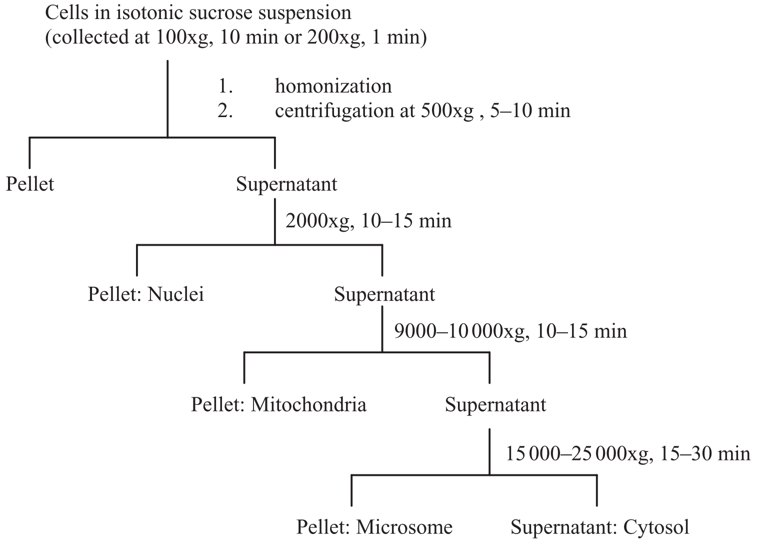

Cells in isotonic sucrose suspension
(collected at 100xg, 10 min or 200xg, 1 min)
1. homonization
2. centrifugation at 500xg , 5–10 min
Pellet
Supernatant
2000xg, 10–15 min
Pellet: Nuclei
Supernatant
9000–10 000xg, 10–15 min
Pellet: Mitochondria
Supernatant
15 000–25 000xg, 15–30 min
Pellet: Microsome
Supernatant: Cytosol

Figure 3.1 Separation of cellular components by differential centrifugation

fractionation of major subcellular components from animal tissues by differential centrifugation. These guidelines are also applicable to plant materials except that chloroplasts can be fractionated from plant leaves by centrifugation between 800 × g for 5 min (supernatant) and 1800 × g for 5 min (pellet).

Each cellular component has its unique physiological function(s)/metabolic activity. Certain enzymes are found to be present either exclusively or predominantly in certain subcellular organelles and can be used as enzyme markers for the characterization of these cellular components. Compartmentalization of enzymes in subcellular organelles is summarized in Table 3.2.

Biochemists generally employed three purification methods:

1. bulk treatment;
2. chromatography; and
3. migratory separation in electric field.

TABLE 3.2 Subcellular localization of enzymes/groups of enzymes

Nucleus	Mitochondrion	Endoplasmic reticulum	Lysosomes	Cytosol
Soluble space:	Matrix:	Smooth ER:	Glycosylic:	Carbohydrate:
Glycolytic Ez	TCAcycle Ez	Cholesterol	β-glucuronidase	Glycogen S
PPP Ez	Fa β-oxid Ez	Biosynthetic	β-NAcHAD	Phosphorylase
Lactate DH	Pyruvate C	Ez	HyaluronoGlc	Glycolysis Ez$^{\#a}$
Malate DH	PEP carboxyK	Steroid	AD	PPP Ez$^{\#b}$
*iso*Citrate DH	CarbamoylP S	hydroxylation	Lysozyme	FDPase
Arginase	Glu DH	Ez	Neuraminidase	PEP carboxyK
Chromatin:	Inner	Fa elongation	Proteolytic:	Malate DH
DNA NTF$^{\#}$	membrane:	Ez	Cathepsins	Lactate DH$^{\#}$
RNA NTF$^{\#}$	Succinate DH$^{\#}$	Carnitine	Elastase	*iso*Citrate DH
NMN	NADH DH	acylTF	Collagenase	Citrate L
adenylylTF	ATPase	GlycerolP	Nucleolytic:	Protein:
NTPase	3-OHbutyrate	acylTF	DNase II	Asp amino TF
Nucleolus:	DH	Drug	RNase II$^{\#}$	Ala aminoTF
RNA NTF	Glycerol3P DH	metabolizing	Lipolytic:	Arginase
RNA Methyl	HK	Ez	Phospholipases	Arginosuccinat
TF	Cyt c oxidase$^{\#}$	Rough ER:	A's	e L
Membrane:		Protein	Cholesterol esterase	Arginosuccinat
G6Pase	Intermembrane:	synthesis	Others:	e S
Acid Pase	Adenylate K	ATPase	Acid Pase$^{\#}$	Aacyl-tRNA S
	NMP K	NDPase	Aryl sulfatase	Nucleic acid:
	NDP K	G6Pase$^{\#}$		Nucleoside K
		NADPH-Cyt		Nucleotide K
	Outer membrane:	reductase		Lipid:
	NADH DH	NDP		Aceyl CoA C
	Cyt b_5 reductase	GlycosylTF		Fa S
	Amine oxidase	Cholesterol		Glycerol3P DH
	Acyl CoA S	acylTF		
	GlyceroP acylTF			
	CholinoPTF			
	Adenylate K			
	HK			
	Phospholipase A_2			

Note: 1. Abbreviations used are: AD, aminidase; C, carboxylase; DH, dehydrogenase; Ez, enzyme/enzymes; ER, endoplasmic reticulum; Fa, fatty acid; H, hexo-/hexose; L, lyase; P, phosphatate; Pase, phosphatase; K, kinase; PEP, phosphoenolpyruvate; PPP, pentose phosphate pathway; NM/D/TP, nucleoside mono/di/triphosphate; NTF, nucleotidyltransferase; S, synthese/synthetase; TF, transferase.

2. Marker enzymes commonly used in subcellular fractionation are denoted with #. The marker enzymes for glycolysis (#a) and pentose phosphate pathway (#b) are 6-phosphofructokinase and G6P dehydrogenase respectively.

Bulk treatment involves salt fractionation, organic solvent fractionation, pH precipitation and adsorbent treatment. The solubility of charged substances such as proteins and nucleic acids generally increases with the salt concentration (salting in) at low ionic strength, but decreases with the salt concentration (salting out) at high ionic strength. Glycans can be extracted with neutral salt (e.g. $CaCl_2$, $MgCl_2$) solutions. Salting out is the basis of the salt (e.g. ammonium sulfate) fractionation used in protein purification. Water miscible organic solvents such as ethanol, n-butanol and acetone are used to isolated nucleic acids, fractionate proteins and purify polysaccharides. Oligomers or polymers (DP >20) of biomolecules can be precipitated with two volumes of alcohol. This precipitation removes salts and low molecular weight materials such as monosaccharides, amino acids, peptides and nucleotides. The precipitate can then be dissolved in water. The charged nucleic acids and proteins can be separated from neutral polysaccharides by ion-exchange chromatography on DEAE-cellulose or CM-cellulose. The solubility of a protein in a dilute salt solution is at its minimum near its isoelectric pH (p*I*). This property is exploited in the isoelectric precipitation of proteins in which the pH of a protein mixture is adjusted to the p*I* of the protein to be isolated so as to minimize its solubility. The most widely used adsorbents are charcoal, aluminum hydroxide gel and tricalcium phosphate (hydreoxyapetite) gel. Selective adsorption is carried out in two ways; negative adsorption in which the contaminants having high affinity are removed by the adsorbent, and positive adsorption in which the desired biomacromolecule is preferentially adsorbed and subsequently desorbed from the absorbent. The bulk method achieves quick removal of some contaminants with a concomitant concentration of the bulk volume. This method is often employed in an early step of purification procedures. A small-scale separation/purification of DNA, RNA and proteins can be accomplished by the use of CsCl density gradient centrifugation. For example, in CsCl density gradient of 1.6–1.8 g/mL, the DNA bands near the center of the centrifuge tube, RNA pellets to the bottom, and proteins float near the top. The chromatographic and electrophoretic techniques, which form the core procedures of biomacromolecular purification, will be discussed in the next section.

The characteristics of biomacromolecules such as solubility, ionic charge, polarity, molecular size, and specificity are utilized in the design of various purification procedures, as shown in Table 3.3.

3.2 PURIFICATION: CHROMATOGRAPHY

The most common form of chromatography (Edward, 1970, Miller, 2005) for purification purposes is liquid chromatography (LC), in which a mixture of substances (solutes) to be fractionated is dissolved in a liquid known as the mobile phase. The resultant solution is percolated through a column packed with a porous solid matrix known as the stationary phase. The interaction of the solutes with the stationary phase results in the differential migration and separation of each component into bands/fractions of pure substances.

Liquid chromatography, or column chromatography, is an ideal technique for purification and analysis of biomacromolecules (Millner, 1999; Schoenmakers, 1986). The position or retention time of a chromatographic peak is governed mainly by the fundamental thermodynamics of solute partitioning between mobile and stationary phases. In a chromatographic operation, capacity is a measure of the amount of solute that can be adsorbed from solution onto a unit volume or weight of the stationary phase, while resolution is a measure of the degree of separation of a desired solute molecule from contaminants. The related parameters are the capacity factor (k) and the resolution (R).

TABLE 3.3 Purification procedures based on biomacromolecular characteristics

Characteristics	Procedure
Solubility	Salt fractionation Organic solvent fractionation Heat treatment
Ionic charge	pH precipitation Ion exchange chromatography Gel electrophoresis Isoelectric focusing
Polarity	Organic solvent fractionation Absorbent treatment Adsorption chromatography Hydrophobic interaction chromatography
Molecular size	Dialysis and ultrafiltration Ultracentrifugation Gel filtration/permeation chromatography Gel electrophoresis
Specificity	Affinity chromatography

The capacity factor, the time required or the number or volume to elute a particular solute, is estimated by

$$k = t_r/t_0 - 1 \text{ or } k = v_r/v_0 - 1,$$

where t_r, is retention time and v_r is elution volume of the peak measured at the peak maximum. and t_0 is dead time and v_0 is void volume of the column. The capacity factor is related to the selectivity factor (α) by

$$\alpha = k_1/k_2 \text{ and retention factor by } k/(1 + k)$$

The resolution is calculated according to

$$\mathrm{R} = (t_{r1} - t_{r2})/\{2(W_1 - W_2) \text{ or } \mathrm{R} = (v_{r1} - v_{r2})/\{2(W_1 - W_2)$$

where t_{r1} and t_{r2} are retention times, and v_{r1} and v_{r2} are elution volumes. W_1 and W_2 are full widths at the peak base of the first and second peaks respectively. The resolution is related to selectivity factor, efficient factor and retention factor by

$$R = 1/4(\alpha - 1)(N^{1/2})\mathrm{k}/(1 + \mathrm{k})$$

where N is efficiency factor, i.e. $N = 16(t_r/W)^2 = 5.54(t_rW_{h/2})^2$.

Various liquid chromatographic methods are classified according to the stationary phases and their modes of interaction, as represented in Table 3.4.

Gel filtration/permeation chromatography (also known as molecular exclusion chromatography) is a form of partition chromatography in which the solute molecules are partitioned between solvent and a stationary phase of defined porosity without an attractive interaction between the two phases. Gel filtration generally refers to aqueous systems while gel permeation is used in nonaqueous systems. The technique is normally used for the separation of biomacromolecules on the basis of size. Solutes are eluted in the order of decreasing molecular size. Gel filtration chromatography is not used as the first step in

TABLE 3.4 Characteristics of chromatographic methods

Chromatographic method	Solute property exploited	Relative capacity	Resolution	Average yield
Gel filtration chromatography	Size	Low	Low	High
Adsorption chromatography	Polarity	Medium-high	Medium	Medium
Partition chromatography	Polarity	Medium-high	Medium	Medium
Ion exchange chromatography	Charge	High	Medium	Medium
Chromatofocusing	Charge	Low	High	Medium
Affinity chromatography	Biospecificity	Medium	Very high	Low

Notes: The solid support of the stationary phase is known as the matrix, which includes cellulose, dextran, agarose, polyacrylamide, silica, polystyrene, polyvinyl, polyether and resins. The matrix of differing porosity and stability varies in the degree of cross-linkages. It is modified to confer specific properties the chemical attachment of various functional groups.

biomacromolecular purification because of its low capacity, but is often used after salt fractionation or pH treatment.

The volume of eluent between the point of injection and the peak maximum is known as the elution volume (V_e), which is used to characterize the behavior of the solute molecule and is dependent on the fraction of the stationary phase available for diffusion by

$$K_{av} = (V_e - V_0)/(V_t - V_0)$$

where, K_{av} = apparent distribution coefficient describing the fraction of stationary gel volume available for diffusion, V_0 = void (exclusion) volume, i.e. volume occupied by the solvent, and V_t = total bed volume.

Dextran (for range 0.05–600 kDa), polyacrylamide (for range 0.1–400 kDa) and agarose (for range 10–40 000 kDa) of various porosities are common matrices used in gel filtration chromatographic separation/fractionation of biomacromolecules, particularly polysaccharides.

Adsorption chromatography (liquid–solid adsorption chromatography) can be considered as a competition between the solute and solvent molecules for adsorption sites on the solid surface of adsorbent to effect separation. In normal phase or liquid–solid chromatography, relatively nonpolar organic eluents are used with the polar adsorbent to separate solutes in order of increasing polarity. In reverse phase chromatography, solute retention is mainly due to hydrophobic interactions between the solutes and the hydrophobic surface of adsorbent. The polar mobile phase is used to elute solutes in order of decreasing polarity. The hydrophilic–hydrophobic interaction is influenced by certain ions present according to Hofmeister series:

Anions: HPO_4^{2-}, SO_4^{2-}, $CH_3CO_2^-$, Cl^-, Br^-, NO_3^-, ClO_4^-, I^-, SCN^-
◄—— Increasing salting-out effect ——
Cations: NH_4, Rb^+, K^+, Na^+, Cs^+, Li^+, Mg^{2+}, Ca^{2+}, Ba^{2+}
—— Increasing salting-in effect ——►

Salting-out ions decrease the availability of water molecules in solutions and enhance hydrophobic interaction. By contrast, salting-in or chaotropic ions prevent non-ionic interaction by ordering water structure, and their use in elution from hydrophobic matrix is usually avoided.

Hydroxyapetite (insoluble form of calcium phosphate with empirical formula of $Ca_3(PO_4)_3OH$) is particularly useful for chromatographic purification/fractionation of DNA. Double-stranded DNA binds to hydroxyapetite more tightly than RNA and proteins, which can be released with a low concentration of phosphate buffer. An increase in the concentrations of phosphate buffer elutes the single-stranded DNA, which is followed by the double-stranded DNA at much higher phosphate concentrations.

In partition chromatography (liquid–liquid partition chromatography), a partition packing consists of a liquid phase coated on an inert solid. The separation is effected by the interaction of the solute between the mobile phase and the liquid stationary phase.

Ion-exchange chromatography separates compounds on the basis of their molecular charges. Compounds capable of ionization, particularly zwitterionic compounds, separate well on ion-exchange column. The separation proceeds because ions of opposite charge are retained to different extents. The resolution is influenced by:

- the pH of the eluent, which affects the selectivity; and
- the ionic strength of the buffer, which mainly affects the retention.

Ion-exchange chromatography is generally used in an early stage of biomacromolecular purification. It is used to separate neutral biomacromolecules (e.g. neutral polysaccharides) from charged biomacromolecules such as nucleic acids versus proteins.

For an effective use of ion exchange in biomacromolecular purification, the stationary phase must be capable of binding either positively or negatively charged molecules. Thus ion-exchange matrices are derivatized with positively charged groups for the adsorption of anionic biomacromolecules (termed anion exchangers) or negatively charged groups for the adsorption of cationic substances (termed cation exchangers), as shown in Table 3.5.

Associated with both stationary phase and charged groups on biomacromolecules are counter ions, which screen the ionogenic groups of the exchangers affecting the binding between ion-exchanger and biomolecules. The counter ions can be arranged in an activity series that represents their strength of interaction with their respective ionogenic groups at equal concentration:

Cations: $Ag^+ > Cs^+ > Rb^+ > K^+ \geq NH_4^+ > Na^+ > H^+ > Li^+$

Anions: $I^- > NO_3^- > H_2PO_4^- > CN^- > Cl^- > HCO_3^- > HCO_2^- > CH_3CO_2^- > OH^- > F^-$

TABLE 3.5 Common ionogenic groups of exchangers

	Abbrev.	Name	Formula
Anion exchanger			
Weak	AE	Aminoethyl-	$\text{-(CH}_2)_2\text{NH3}^+$
	DEAE	Diethylaminoethyl-	$\text{-(CH}_2)_2\text{NH(C}_2\text{H}_5)_2$
Strong	QAE	Diethyl-2-hydroxypropylaminoethyl-	$\text{-(CH}_2)_2\text{N}^+\text{(C}_2\text{H}_5)_2\text{CH}_2\text{CH(OH)CH}_3$
	TAM	Triethylaminoethyl-	$\text{-CH}_2\text{N}^+\text{(CH}_3)_3$
	TEAE	Triethylaminoethyl-	$\text{-(CH}_2)_2\text{N}^+\text{(C}_2\text{H}_5)_3$
Cation exchanger			
Weak	C	Carboxy-	-COO^-
	CM	Carboxymethyl-	$\text{-CH}_2\text{COO}^-$
Strong	S	Sulfo-	-SO_3^-
	SM	Sulfomethyl-	$\text{-CH}_2\ \text{SO}_3^-$
	SP	Sulfopropyl-	$\text{-C}_3\text{H}_6\ \text{SO}_3^-$

Consequently Na^+ would replace H^+ as the counter ion for a cation exchanger, whilst Cl^- would replace OH^- as the counter ion for an anion exchanger.

Chromatofocusing combines ion-exchange column chromatography with isoelectric focusing separation. The charged group on an ion-exchange resin has a buffering action at a particular pH, therefore it will form a pH gradient if eluted with a second buffer at a different pH. Chromatofocusing is applied to purify proteins based on the difference in their isoelectric points (p*I*). If proteins are bound to the ion exchange resin, they elute as the generated gradient reaches their p*I*s. When a protein and eluting buffer first enter the column, the protein may travel down the column if the *pI* is greater than the initial pH. As the eluting buffer travels through the column, its pH increases until it is greater than the p*I* of the migrating protein. The protein reverses its formal charge and adheres to the matrix. The subsequent elution with a pH gradient developed by the second buffer with lower pH decreases the pH below the p*I* of the protein, which is then released from the anion exchanger, and eventually eluted from the column at its isoelectric pH.

Affinity chromatography is a type of absorption chromatography in which the molecule to be purified is specifically and reversibly adsorbed by a complementary binding ligand immobilized on the matrix (Mohr and Pommerening, 1985; Turkova, 1978). It is applied to purify biomolecules on the basis of their biological function or specific chemical structure. The highly specific method of purification utilizes the specific reversible interactions between biomolecules. Purification is often in the order of several thousand-fold and recoveries of the purified biomolecule are generally high.

The optimal situation for an affinity chromatography is achieved when only the specifically retarded biomacromolecule either remains associated with the immobilized specific ligand until elution is effected by changing the chromatographic condition or is retarded sufficiently to achieve a complete resolution from void volume nonspecific biomolecules. An important parameter in determining the retardation of a biomacromolecule on an affinity matrix is the dissociation constant, K_a for the interaction of the biomacromolecule, P and the ligand, L:

$$\mathrm{K_1 = [P][L]/[PL] = [P_0 - PL][L_0 - PL]/[PL] \approx [P_0 - PL][L_0]/[PL]}$$

where P_0 and L_0 are the initial concentration of biomacromolecule and the concentration of the immobilized ligand respectively for the reaction $P + L \leftrightarrow PL$, under which $P_0 \gg L_0$.

We can now define the chromatographic distribution coefficient, K_d as

$$K_d = [\mathrm{PL}]/[\mathrm{P_0 - PL}] = L_0/K_1$$

and the elution volume, V_e of the biomacromolecule, which interacts specifically with the ligand, is defined as

$$V_e = V_0 + K_d V_0$$

where V_0 is the void volume of the matrix. Therefore the retardation of the specifically interacting biomacromolecule is

$$V_e/V_0 = 1 + L_0/K_1$$

which is determined by the concentration of immobilized ligand and the dissociation of the biomacromolecule–ligand complex.

The ideal matrix for use in affinity chromatography should be highly porous yet rigid, hydrophilic, uniform in size and chemically stable to conditions used to immobilize the appropriate ligand. Agarose is the most commonly used matrix material. A spacer molecule (generally corresponding to 6–8 methylene groups) is often employed to distance the ligand from the matrix to facilitate the interaction between biomacromolecule and ligand. The ligand should be specific, stable and bind biomacromolecule of interest reversibly with high affinity. For substantial adsorption of the biomacromolecule from solution (e.g. $[PL]:[P_0]$ $(L_0:K_l) = 95:5$), the value of K_l must be about two orders of magnitude less than the concentration of immobilized ligand.

The base pairings between A–T(U) and G–C are exploited in the purification of nucleic acids, such as purification of specific genes with immobilized oligonucleotides. An interesting example of applying hybridization between DNA and mRNA is the purification of eukaryotic mRNA (having a polyA sequence at 3′ ends) with polyU affinity chromatography.

Suitable pairs of protein and ligand combination for affinity chromatography are antigen-antibody, hormone-receptor, glycoprotein-lectin or enzyme-substrate/cofactor/effector (Table 3.6). The affinity of lectins for specific carbohydrate moieties (Lis and Sharon, 1973; Barondes, 1981) has made them particularly useful for the purification of distinct groups of glycans and glycoproteins (Table 3.7).

The fundamental principles for the conventional LC and high performance (high-pressure) chromatography (HPLC) remain the same. HPLC separation/purification of biomacromolecules (Aguilar, 2004; Millner, 1999) differs from the conventional LC in the following ways:

- The sorbents are of much greater mechanical strength.
- Sorbent particle size has been decreased about 10-fold to enhance adsorption-desorption kinetics and diminish bandspreading.
- The columns are operated at 10–60 times higher mobile-phase velocity by the use of high-pressure metering pumps (Table 3.8).

This enables the samples to be introduced as narrow bands, higher and rapid resolution of the mixtures and an incorporation of automated analytical method to facilitate purification processes. An improvement in the speed of separation for biomacromolecules can be achieved by the use of superficially porous support packed into microbore columns

TABLE 3.6 Examples of ligands suitable for affinity chromatographic purification of proteins

Ligand	Protein to be purified
Nucleotides	Nucleotide binding/requiring proteins/enzymes
Sugars	Lectins, glycosidases, sugar transporter proteins
Fatty acids	Fatty acid binding proteins, albumin
Steroids	Steroid hormone receptors
Amino acids	Amino acid binding/metabolizing proteins/enzymes
Antigens	Monoclonal/polyclonal antibodies
Monoclonal antibodies	Antigens
Protease inhibitors	Proteases
Lectins	glycoproteins
Avidin	Biotin-containing proteins
Heparin	Coagulation factors
Phenylboranate	Glycated proteins, glycoproteins
Alprenolol	β-adrenergic receptor
Lysine	Plasminogen, plasmin
ATP	Kinases
$NAD(P)^+$	Dehydrogenases

TABLE 3.7 Specificity of some lectins used in glycan affinity chromatography

Lectin	Abbrev.	Mol. wt.	Specificity/binding	Sugar for elution
Concanavalin A (jack bean)	Con A	55 000	α-D-Man, α-D-Glc	10 mM α-Me-Glc and 100 mM α-Me-Man
Lentil lectin	LCA	49 000	α-D-Man, α-D-Glc	100 mM α-Me-Man
Soybean lectin	SBA	115 000	D-GalNAc	50 mM GalNAc
Castor bean lectins	RCA$_{60}$	60 000	D-GalNAc	50 mM GalNAc
	RCA$_{120}$	120 000	D-Gal	
Wheatgerm agglutinin	WGA	36 000	(β-D-GlcNAc)$_3$, α-D-NeuNAc	10 mM chitotriose
Peanut agglutinin	PNA	120 000	β-D-Gal(1 → 3)-D-GalNAc	
Pea lectin (garden pea)		49 000	α-D-Man	100 mM α-Me-Man
Phytohaemagglutinin	PHA	128 000	D-GalNAc	50 mM GalNAc
Pokeweed mitogen	PWM	32 000	(D-GlcNAc)$_3$	10 mM chitotriose
Potato aggutinin			(D-GalNAc)$_3$	
Jacalin		40 000	α-D-Gal	100 mM α-Me-Gal
Limulin		400 000	α-D-NeuNAc	
Sambucus nigra agglutinin	SNA		α-D-NeuNAc(2 → 6) D-GalNAc	100 mM Lactose
Ulex europeus agglutinin	UEA1	170 000	α-L-Fucose	25 mM Fucose
Helix pomatia lectin	HPA	79 000	α-D-Gal, α-D-GalNAc	50 mM GalNAc

TABLE 3.8 Comparison of low *versus* high pressure chromatography

Characteristic	Low pressure (conventional)	High pressure
Particle size (μm) of matrix	~100	~10
Flow rate (ml cm^{-2} h^{-1})	10–30	100–600
Operating pressure (atm)	<5	>50
Separation time (hr)	Up to 48	1–3
Sample volume	mL	μL to mL
Resolution	Good	Excellent
Common matrix	Cellulose, dextran, agarose, polyacrylate, polystyrene, polyether, silica	Silica, resins

(Kirkland *et al.*, 2002). For example, the superficial porous support (5 μm) with solid silica core and 0.25 μm of superficial porous ring packed into 2.1 mm i.d. column permits the separation of biomacromolecules at flow rate of 1–2 mL/min.

3.3 PURIFICATION: ELECTROPHORESIS

The transport of particles by an electrical field is called electrophoresis. The electrophoretic mobility (U), which is the ratio of velocity of moving particles (v) to the strength (potential) of the electrical field (E), is related to the number of charges (Z) and the magnitude of the particle charge (e) by the expression:

$$U = v/E = Ze/f' = Ze/(6\pi\eta r)$$

where f is the frictional coefficient that is related to the particle radius, r (Stokes' radius) and the solvent viscosity according to the Stokes' law by $f = 6\pi\eta r$ for spherical particles.

Thus the mobility of particles with zero charge would be zero. Charged biomacromolecules, especially nucleic acids at pH above 2.4 (Rickwood and Hames, 1982), proteins at pH other than their isoelectric points (Hames and Rickwood, 1981), and acidic polysaccharides (e.g. chondroitin sulfate, heparin, hyaluronic acid) migrate in an electric field with the rates of migration dependent upon their charge densities. Therefore electrophoresis is an ideal technique to resolve and separate the individual component of charged biomacromolecular mixtures (Andrews, 1986; Westermeier, 2001).

Zone electrophoresis is the separation of charged molecules in a supporting medium, resulting in the migration of charged species in discrete zones (Andrews, 1986) to minimize the effect of convection and diffusion. Various gels (such as agar, agarose and polyacrylamide) used as the supporting media may also exert a molecular sieving effect. This allows gel electrophoresis to separate charged biomolecules according to their mobilities (size, shape and charge), the applied current and the resistance of the medium. Polyacrylamide gels can be prepared with pores of the same order of size as protein/oligonucleotide molecules and so are effective in the fractionation of proteins and oligonucleotides (Hames and Rickwood, 1981). Also, agarose gels are used to separate larger molecules or complexes such as certain nucleic acids and nucleoproteins (Rickwood and Hames, 1982).

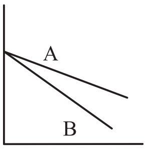
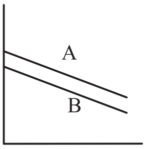
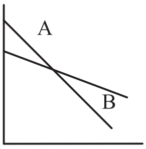
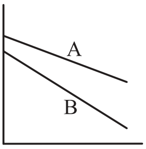
The gel concentration required for polyacrylamide gel electrophoresis (PAGE) to achieve optimal resolution of two proteins (or nucleic acids) can be determined by measuring the relative mobility of each protein in a series of gels of different acrylamide concentrations to construct a Ferguson plot ($\log R_f$ versus C) according to

$$\log R_f = Y_0 + \mathrm{K}_r C$$

where C is the gel concentration (e.g. percent total monomer of acrylamide + N,N′-methylene bisacrylamide in polyacrylamide gel related to the gel pore size) and R_f is the relative mobility of the charged macromolecule (i.e. distance migrated by macromolecule/distance migrated by marker macromolecule or tracking dye). Each Ferguson plot is characterized by its slope (K_r, retardation coefficient) and ordinate intercept (Y_o, i.e. $\log R_0$, the relative mobility in free electrophoresis). The retardation coefficient is related to the molecular size of the protein and the relative mobility is a measure of the mobility of the macromolecule in solution that is related to its charge. Thus the choice of the gel concentration greatly affects the separation/purification of charged biomacromolecules.

Four cases are possible for the gel electrophoretic separation/purification of proteins that differ in the charge and/or size (Table 3.9). Case 1 (equal charge but different size) approximates sodium dodecylsulfate (SDS)-PAGE, while case 2 (different charge but

TABLE 3.9 Effect of gel pore size on separation of charged biomacromolecules

Case	1	2	3	4
Property:	Equal charge but different size	Different charge but equal size	Different charge and size	Different charge and size
Charge	A = B	A > B	A > B	A > B
Size	A > B	A = B	A > B	A < B
$\log R_f$ vs. C plot				

equal size) typifies isozymes. Cases 3 and 4 represent separation of proteins with differences in both charge densities and sizes. Case 3 also characterizes the electrophoretic separation of nucleic acids. The electrophoretic mobility of the DNA fragments is dependent on fragment size and fairly independent of base composition or sequence. In the pH 5.0–8.0 range, each nucleotide residue bears a charge of −1, and the net negative charge of RNA and single-stranded DNA molecules are nearly equal to their chain length (the net negative charge of the double-stranded DNA is double of the chain length). Thus for a given gel concentration, there will usually be some chain length range in which logarithm of the number of nucleotide residues per chain and relative mobility are approximately linearly related according to

$$U = \mathrm{a} - \mathrm{b}\log N$$

where U is the mobility expressed in $\mathrm{cm}^2/\mathrm{V\,sec}$, N is the chain length (i.e. number of nucleotide residues/bases), and a and b are constants depending on gel composition and temperature. Therefore gel electrophoresis is a convenient method for separation/purification of nucleic acids according to their molecular sizes (chain lengths). For sequence analysis of oligodeoxynucleotide/DNA fragments, polyacrylamide gels can be used for fragments between 6 base pairs (bp) (20% acrylamide) and 1 kbp (3% acrylamide), while agarose gels are used to analyze double-stranded DNA fragments from 70 bp (3% agarose gel) to 800 kpb (0.1% agarose gel). Since the migration of DNA in a gel depends on molecular dimension of DNA, circular DNA migrates differently from linear DNA and supercoiled topoisomers. Denaturing gel systems are available for the analysis of single stranded DNA fragments. Thus a high resolution PAGE, such as a thin (0.035 mm) polyacrylamide slab containing 7 M urea, is used for the sequence analysis of DNA fragments.

Under appropriate conditions, all reduced polypeptides bind the same amount of sodium dodecylsulfate (SDS), i.e. 1.4 g SDS/g polypeptide. Furthermore, the reduced polypeptide-SDS complexes form rod-like particles with lengths proportional to the molecular weight of the polypeptides. This forms the basis for separation/analysis of proteins according to their chain weight by SDS-PAGE (Weber et al., 1972). Thus there exists an approximately linear relationship between the logarithm of the protein subunit weight and mobility for each particular SDS-gel and concentration. As a rule, the linear relationship exists only over a limited range of molecular weight (MW). For example, 15% acrylamide for MW of 12 000–45 000, 10% acrylamide for MW of 15 000–70 000, and 5% acrylamide for MW of 25 000–200 000. The electrophoretic separation of large MW DNA (>100 kbp) is hampered by the orientation of the fully elongated chain that becomes trapped in the gel network. Because large DNA molecules tend to become trapped in this gel network and are unable to migrate, an alternative technique known as pulsed field gel electrophoresis (PFGE) is conceived to separate large DNA molecules ($>10^6$ bp). This involves application of a periodic, brief reversal or change in direction of the field, allowing the elongated DNA molecules to become disentangled and resume migration (Birren and Lai, 1993).

Capillary electrophoresis has been developed for the separation of nucleic acids and proteins including recombinant proteins (Strege and Lagu, 2004). It is performed in a typical capillary column with a dimension of 50 μm (i.d). × 30–100 cm (length), and the injection volume is limited to <20 nL. The technique offers a sensitive and high resolution of charged biomacromolecules. However, the limited sample injection volume in capillary electrophoresis limits the detection of minor components in complex mixtures, and suitable pre-concentration is often needed.

Electrofocusing (EF) or isoelectric focusing (IEF) is a charge fractionation technique that separates molecules predominantly by the difference in their net charge, not by size

(Allen, Saravis and Maurer, 1984). Thus EF can be considered as an electrophoretic technique by which amphoteric compounds are fractionated according to their isoelectric points (p*I*s) along a continuous pH gradient maintained by ampholyte buffers (ampholytes are mixtures of relatively small, multicharged amphoteric molecules with closely spaced p*I* values and high conductivity). The pH gradients can be designed to achieved optimal resolution of amphoteric components by 'buffer electrofocusing' (Nguyen and Chramback, 1980) or by 'constituent displacement' via pH of the catholyte/anolyte (McCormick, Miles and Chrambach, 1976). This is contrary to zone electrophoresis, where the constant pH of the separation medium establishes a constant charge density at the molecule and causes it to migrate with constant mobility. The charge of an amphoteric macromolecule in EF decreases according to its titration curve, as it moves along the pH gradient approaching its equilibrium position at p*I*. Here the molecule comes to a stop and condenses (focuses) into a sharp band in the pH gradient at its characteristic p*I* value. Proteins with difference in p*I*s of 0.02 pH units can be resolved by EF.

Two-dimensional electrophoresis (2DE) describes a variety of methods employing separation of biomolecules in two dimensions, of which two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) is specifically applied to the separation and analysis of nucleic acids (Burckhardt and Birnstiel, 1978) and proteins (Dunbar, 1987). The relative order of migration of the components of the mixture should be very different in two dimensions of an optimal 2D-PAGE. The main factors that can be altered to obtain a different migration order are gel concentration, pH and the presence or absence of a denaturing agent (e.g. SDS, urea). A wide selection of protein detection scheme is available for the visualization of proteins separated by 2DE (Rabilloud, 2000). Some of the combinations in 2D-PAGE are shown in Table 3.10.

The introduction of the immobilized pH gradient (IPG strip) greatly enhances the potential of 2DE (Righetti and Bossi, 1997; Fichmann, 1999; Görg *et al.*, 2000). This is

TABLE 3.10 Examples of gel combinations in 2DE

		First dimension			Second dimension		
Biomacromolecule		Gel conc.	pH	Denaturant	Gel conc.	pH	Denaturant
DNA	Rest. frag.	8%	7.7	0	4%	8.3	0
	Rest. frag.	1% agarose	8.0	0	4%	8.0	7 M urea
RNA	tRNA	16%	8.0	6 M urea	16%	8.0	0
	mRNA	6%	8.0	0	6%	8.0	5 M urea
	vRNA frag.	10%	3.3	6 M urea	20%	8.0	0
Proteins		4%	0.4% amph. (pH 3.5–10)	8 M urea gradient	10–16% gel	8.8	0.1% SDS
Proteins		3%	2% amph. (pH 3.5–10)	8 M urea gradient	5–20% gel	8.8	1% SDS
Membrane proteins		4%	2% amph. (pH 3.5–10)	4 M urea	5–20% gel gradient	8.8	4 M urea 0.1% SDS
Ribosmonal proteins		4%	8.6	6 M urea	18%	4.0	6 M urea
Histones		15%	0.9 M HOAc	2.5 M urea	15%	0.9 M HOAc 1% Triton	2.5 M urea

Note: Gel concentration refers to acrylamide concentration unless otherwise stated. Abbreviations used are: rest., restriction; frag., fragments; amph., ampholyte; SDS, sodium dodecylsulfate and HOAc, acetic acid.

achieved by the co-polymerization with acrylamide of a set of monomers that carry ampholyte functionality, $CH_2{=}CHCONHR$ with R containing either a carboxyl or tertiary amino group to form buffers with different pK values. Therefore by changing the concentration of the different monomers along the strip, pH gradients are covalently immobilized into the gel strips. Alternatively, ranges of ready-made IPG strips are available commercially as Immobiline DryStrip (Amersham Pharmacia) or ReadySrip (Bio-Rad). Strips of 18 or 24 cm are usually employed for high-resolution separations, while shorter strips (4, 7 or 11 cm) are used for rapid screening applications. The IPG strip is re-swelled in sample solution (containing 8 M urea or 2 thiourea and 5 M urea plus 4% nonionic or zwitterionic detergent such as Triton X-100 or 3[(cholamidoprpyl)dimethylamino]-1-propane sulfonate, 15 mM DTT of the appropriate pH range) prior to focusing. This ensures the uniform distribution of the sample mixture across the whole IPG. Once the strip is re-swelled, an electric field is gradually applied across the IPG. Various integrated instruments, IPGphor (Amersham Pharmacia), Proteam IEF cell (Bio-Rad) and IPGpHaser (Genomic Solution) are available for 2DE operation. The biomacromolecules that are positively/negatively charged move toward the cathode/anode and encounter an increasing/decreasing pH until reaching their p*I*. Thus every biomacromolecule in the mixture is separated and concentrated/focused at their respective p*I*. The gel for the second dimension is cast by polymerizing an acrylamide solution between two glass plates in a slab. After proper equilibration (IEF gels are equilibrated to allow the separated biomacromolecules to interact fully with SDS so that they may migrate properly during SDS-PAGE), the first dimension IPG strip is applied to the second dimension where they are separated according to their molecular weight. Polyacrylamide gels of either single concentrations or gradient concentrations may be used.

2DE has become one of the main methods for the separation/characterization of peptides and proteins for analytical and preparative purposes (Görg *et al*., 2000). It is generally carried out in the first dimension according to their isoelectric points using EF with carrier ampholytes, followed by separation in the second dimension according to their molecular weight using SDS-PAGE. One of the common uses of 2D-PAGE is the rapid purification of a small quantity of protein, which can be cut from the gel for direct sequence determination, and also for antibody purification. The use of 2D-PAGE with silver staining provides one of the best methods to estimate protein purity.

3.4 CHARACTERIZATION: GENERAL

3.4.1 Purity

Prior to characterization, the purity (homogeneity) of a biomacromolecule has to be established. The operational criterion for establishing purity takes the form of the method of exhaustion: the demonstration by all available methods that the sample of interest consists of only one component. Table 3.11 lists the common methods employed to assess the purity of a biomacromolecule preparation (Sheehan, 2000). The relationships among various analytical methods to assess homogeneity and structural integrity of a biomacromolecule are presented in Figure 3.2.

3.4.2 Molecular weight

Biomacromolecules can be characterized by many numerical values. One of the most common and important is the molecular weight (M) as represented in Table 3.12. It is an

TABLE 3.11 Methods for establishing purity

Method	Biomacromolecule	Property assessed	Sample range
Chromatography			
Gel filtration	All	Size	µg
Adsorption	Glycans	Hydrophilicity	µg
Ion exchange	Nucleic acids, proteins	Net charge	µg
Affinity	All	Specific binding	ng
Electrophoresis			
Native gel	Nucleic acids, proteins	Size/charge	ng
Denaturing gel	Nucleic acids, proteins	Chain length	ng
Isoelectrofocusing	Proteins	Isoelectric pH	ng
Sedimentation	All	Mass, size	variable
Spectrophotometry	Nucleic acids, proteins	Chromophore	µg
Composition	All	Component content, specific chemical group	variable
Biological activity	Proteins	Enzyme activity	variable

Note: The method is applicable to the majority of native biomacromolecules in that class.

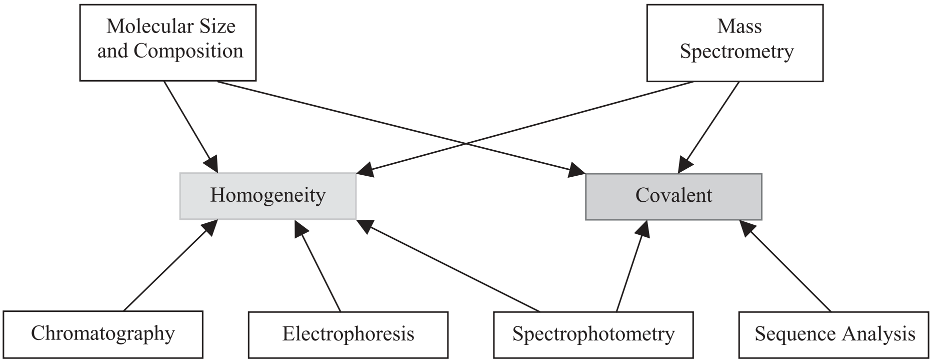


Figure 3.2 Various methods for establishing homogeneity and covalent structure of biomacromolecules

TABLE 3.12 Comparison of M_n *versus* M_w of selected biomacromolecules

Biomacromolecule	M_n	M_w	M_w/M_n
Ovalbumin	45×10^3	46×10^3	1.0
Serum albumin	69×10^3	70×10^3	1.0
Cellulose trinitrate	94×10^3	273×10^3	3.7
Amylopectin	300×10^3	80×10^6	267

Note: M_n and M_w are determined by osmotic pressure and light scattering respectively. The ratio of M_w/M_n differentiates the monodisperse protein versus the polydisperse glycan samples.

easy-to-understand descriptor and perhaps the most often cited characteristic of the molecule. Methods that provide molar quantitation of constituent monomers or covalently attached chemical components can be used to estimate the minimum molecular weight, M_{min} of a biomacromolecule according to

$$M_{min} = m/n$$

in which m is the mass of the biomacromolecule and n in the number of moles of the constituents measured.

If all of the molecules in a biopolymeric sample such as polysaccharides do not have the same value of M, this sample is said to be polydisperse. If all molecules have the same value of M, such as pure sample of DNA and proteins, it is said to be monodisperse. The molecular weight of a polydisperse sample is an average value that can be the number-average molecular weight, M_n or the weight-average molecular weight, M_w, since different experimental techniques for the determination of M yield either M_n (e.g. osmotic pressure) or M_w (e.g. equilibrium centrifugation). M_n is defined as

$$M_n = \Sigma n_i M_i / \Sigma n_i = \Sigma f_i M_i$$

in which n_i is the number of molecules and f_i is the fraction of the total number of molecules having molecular weight M_i. M_w is defined as

$$M_w = \Sigma w_i M_i / \Sigma w_i = \Sigma n_i M_i^2 / \Sigma n_i M_i = \Sigma f_i M_i^2 / \Sigma f_i M_i$$

in which $w_i = n_i M_i$ is the weight of all molecules having molecular weight M_i. Specifically M_n is an average over the number fraction of molecules and M_w is an average over the mass fraction of molecules. Thus M_n is more sensitive to lighter molecules and M_w is more sensitive to heavier molecules in the sample. It follows that a contaminant of low molecular weight makes M_n significantly lower than the true M and has only a small effect on M_w. A high molecular weight contaminant raises M_w with virtually no effect on M_n. For a monodiperse sample, $M_n = M_w = M$.

The methodological approaches to determine molecular weight of biomacromolecules can be classified as:

- chemical methods include analyses of composition and colligative properties;
- transport behaviors include sedimentation, gel filtration chromatography and electrophoresis;
- scattering analyses include X-ray diffraction, Rayleigh light scattering and electron microscopy; and
- mass spectrometry.

Operationally these methods can be categorized as physicochemical, empirical and mass spectrometric methods.

3.4.2.1 Physicochemical methods. The physicochemical methods (though mainly of historical interest) from which molecular weights of most biomacromolecules have been determined earlier include osmotic pressure, light scattering and sedimentation.

Osmotic pressure. Dissolution of a solute reduces the chemical potential of the solvent and results in a number of observable phenomena known as colligative properties, such as boiling point elevation, freezing point depression and osmotic pressure. The determination of the extent of change in these properties provides a measure for the molecular weight of the dissolved solute. Thus the molecular weight of a biomacromolecule can be obtained by measuring osmotic pressure across semi-permeable membrane in a membrane osmometer (Kupke, 1960) according to

$$\lim_{c \to 0} \pi/(RTc) = 1/M + Bc$$

where π is measured osmotic pressure in cm or kPa. R is gas constant (84.8 l-cm deg^{-1} mol^{-1} or 8.314 dm^3 kPa mol^{-1} K^{-1}) and c is the concentration of sample in g dm^{-3}. B is a virial coefficient. Thus a plot of $\pi/(RTc)$ versus c gives an intercept of 1/M (i.e. M = 1/intercept).

Light scattering. The scattering produced by a solution containing particles depends on their weight concentration (c in g/L) and molecular weight (M) and on the angle, θ with respect to the incident beam. But it is symmetrical, with regard to forward and backward scattering, if the scattering particles are small compared to the wavelength of light. For such scattering, known as Rayleigh scattering, the ratio of the intensity (i_θ) of the scattered radiation to the intensity (I_0) of the incident radiation, corrected for the oscillating term, defines the Rayleigh ratio, R_θ for an ideal solution:

$$R_\theta = (i_\theta/I_0)\{r^2/(1 + \cos^2\theta)\} = \{2\pi^2 n_0^2 (dn/dc)^2\}/(N_A\lambda^4)cM = KcM$$

$$\text{i.e. } Kc/R_\theta = 1/M$$

where $K = \{2\pi^2 n_0{}^2 (dn/dc)^2\}/(N_A\lambda^4)$ in which n in the refractive index of the solution, dn/dc is measured change in n with a change in c called the specific refractive index increment, N_A is Avogadro's number and λ is the wavelength of the incident radiation. This expression is expanded for a nonideal solution in a power series about c to

$$Kc/R_\theta = 1/M + 2Bc + \ldots$$

When the scattering molecules (e.g. macromolecules) are of comparable size to the wavelength of the scattered radiation, the equation becomes

$$Kc/R_\theta = \{1 + (16\pi^2 R_G^2 \sin^2\theta)/3\lambda^2\}/M$$

where R_G is the radius of gyration of the scattering molecule. Experimentally, the intensities of scattering are recorded at 90° to the incident radiation at different concentrations, and the equation reduced to

$$\lim_{c \to 0} Kc/R_{90} = 1/M + 2Bc$$

Thus, a plot of Kc/R_θ versus c gives an intercept of 1/M (i.e. M = 1/intercept).

Sedimentation. When macromolecules are subject to a strong centrifugal field, they form a boundary moving outward from the center of rotation known as sedimentation, the rate of which is proportional to the applied centrifugal field and the size of molecules. There are two approaches to sedimentation for the determination of molecular weight of a biomacromolecule. In the sedimentation velocity (operated at ~60 000 rpm), the rate of sedimentation measured as sedimentation coefficient, $s\ (s^{-1}) = (dx/dt)/(\omega^2 x)$ is followed. ω is the angular velocity in radian/sec (2π rpm/60) and x is the distance from the center of rotation for the sample at time, t. The molecular weight (M) is then calculated according to the Svedberg equation:

$$M = (RTs)/[D(1 - v\rho)]$$

in which R is gas constant ($8.314 \times 10^7\,\mathrm{erg\,deg^{-1}\,mol^{-1}}$) and D is diffusion coefficient (in $\mathrm{cm^2\,s^{-1}}$). The partial specific volume, v varies between 0.69 and 0.75 $\mathrm{cm^3/g}$ with an average value of 0.72 $\mathrm{cm^3/g}$ for proteins and ~0.50 $\mathrm{cm^3/g}$ for Na-DNA/RNAs, while ρ and c are the density of the medium and sample concentration at x.

In sedimentation equilibrium (operated at ~8000 rpm), the concentration gradient of the sedimenting macromolecules is measured. The molecular weight of biomacromolecule is evaluated according to

$$\ln c = \mathrm{M}[\omega^2(1 - v\rho)/(2\mathrm{RT})]x^2 + \mathrm{k}$$

where c (mg/L) is the concentration of the macromolecule at a distance, x (cm) from the axis of rotation.

A plot of $\ln c$ versus x^2 gives a straight line with a slope of $\mathrm{M}\{\omega^2(1 - v\rho)]/(2\mathrm{RT})\}$ from which M can be calculated (i.e. $\mathrm{M} = \{\mathrm{slope}\}\{\omega^2(1 - v\rho)]/(2\mathrm{RT})\}$).

3.4.2.2 Empirical methods. Some of the simplest, convenient and rapid approaches to the estimation of the molecular weight of a biomacromolecule are empirical methods. The unknown sample is co-analyzed with reference macromolecules of known molecular weights having approximately the same shapes as the biomacromolecule of interest. The molecular weight of the unknown is estimated from the interpolation of the correlation between analytical data and the molecular weights of the reference macromolecules. Three common approaches based on transport behaviors of biomacromolecules are gel filtration chromatography (Andrews, 1970), SDS-PAGE (Weber *et al.*, 1972) and sucrose/density gradient centrifugation (Martin and Ames, 1961), all of which measure relative mobility of biomacromolecules (Table 3.13).

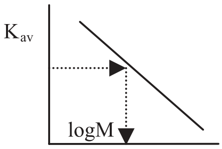
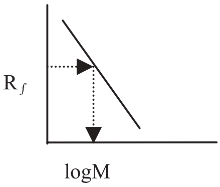
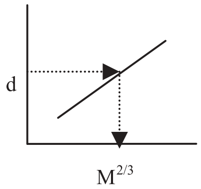
3.4.2.3 Mass spectrometry. Mass spectrometry (MS) provides a rapid and sensitive technique to accurately measure the molecular weight of biomolecules. The mass spectrometers have seven major components: i) a sample inlet; ii) an ion source; iii) a mass analyzer; iv) a detector; v) a vacuum system; vi) instrument control system; and vii) a data system (Figure 3.3).

Variations of instrument components typically used in protein identification and sequencing are:

Sample inlet:	1. Direct probe or stage 2. Capillary column liquid chromatography
Ion source:	1. Electrospray including nanospray and microspray 2. Matrix-assisted laser desorption
Mass analyzer:	1. Quadrupole mass filter 2. Ion trap mass analyzer 3. Time-of-flight (TOF) mass analyzer

The basic processes associated with an MS experiment involve the generation of gas-phase ions derived from an analytical sample (analyte), and the measurement of the mass-to-charge ratio (m/z) of those ions (Biemann, 1992; Watson, 1997; Larson and McEwen, 1998). The challenge to the application of MS to any analytes is the production of gas-phase ions of those species. Fast atom bombardment (FAB) was the first ionization method that made routine MS analysis of polar molecules possible. The development of electrospray ionization (ESI) (Whitehouse *et al.*, 1985) and matrix-assisted laser desorption/ionization (MALDI) (Karas and Hilenkamp, 1988) MS has given biochemists two efficient and robust methods for producing gas-phase ions of oligomeric and polymeric

TABLE 3.13 Empirical methods for molecular weight determination of biomacromolecules

	Gel filtration chromatography	SDS-Polyacrylamide gel electrophoresis	Sucrose gradient centrifugation
Principle	Sieve effect	Electric effect	Centrifugal effect
Empirical relationship	$V_e = \alpha - \beta \log M$ where V_e is elution volume.	$R_f = \alpha - \beta \log M$ where R_f is the relative migration.	$d = \alpha + \beta M^{\gamma}$ where d is the distance from the center of rotation.
Plot	K_{av} vs. logM  	R_f vs. logM  	d vs. $M^{2/3}$ 
Matrix	Dextran, polyacrylamide, agarose	SDS in polyacrylamide	5–20% sucrose gradients
Application	Glycans	Proteins, nucleotides	Proteins, nucleic acids
Reference	Andrew, 1965	Weber *et al.*, 1972	Martin and Ames, 1961

Notes: 1. The use of K_{ave} instead of V_e in gel filtration chromatography generally improve the linear relationship between K_{ave} and M^{γ} (Ui, 1979) where $K_{ave} = (V_e - V_o)/(V_t - V_o)$ in which $V_{e,}$ V_t and V_o are elution, total and void volumes respectively.

2. The use of two gel types in gel filtration chromatography to evaluate the molecular weight of proteins is valuable because the measured values of M will differ for the two gels if the protein violates the K_{av} versus logM relationship. The actual dependence in gel filtration chromatography of proteins is K_{av} *versus* log r (Stokes radius) (Siegel and Monty, 1966).

3. SDS-PAGE yields the subunit molecular weight of proteins.

4. Because the charge-to-mass ratio is almost the same for nucleic acids, the native polyacrylamide gel and agarose gel electrophoresis may be employed to estimate the molecular weight of RNA and single-stranded DNA.

5. For nucleic acids, $s_{20,w} = \alpha + \beta M^{\gamma}$, where $s_{20,w}$ is sedimentation coefficient corrected for water at 20°C. α, β and γ are 2.8, 8.34×10^3 and 0.497 respectively for double-stranded DNA at neutral pH in 1 M NaCl. The S value commonly used in expressing the size of RNA is $S = s_{20,w} \times 10^{13}\,s^{-1}$. The sedimentation can be estimated through linear concentration of sucrose solution (McEwen, 1967). Thus in an isokinetic gradient centrifugation, the distance traveled (d) per unit time can be used in place of $s_{20,w}$ in $d_1/d_2 = (M_1/M_2)^{\gamma}$.

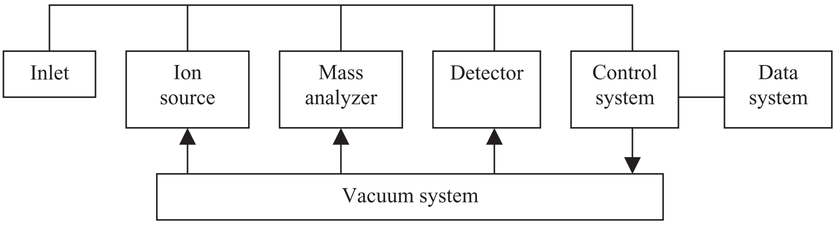


Figure 3.3 A block diagram of a mass spectrometer
The detector, mass analyzer and parts of ion source are maintained under vacuum. The instrument control system monitors and controls all parts of the instrument. Details of the sample inlet, ion source, and mass anlayzer define the type of instrument and the capabilities of that system.

biomolecules. The nature of mass analyzer determines several characteristics of the overall experiment, and the two most important parameters are m/z resolution (mass resolution) and the m/z range of ions that can be measured (mass range). The unit for m/z is the Thomson (abbreviated Th), although many scientists use m/z as a unit-less ratio. It is critical to remember the interchangeable use of ‘mass’ and ‘m/z’, because these values will not be the same for any ion that is multiply charged, particularly when ESI is being used. Two broad classes of ions seen in MS experiments are molecular ions, which contain the entire analyte molecule, and fragment ions, which contain only a portion of the structure. The molecular weight of an analyte can be calculated from the m/z of a molecular ion, if the charge (z) is known, whereas structural information is derived from measuring the m/z of the fragment ions.

In the ESI of biomolecules, an acidic aqueous solution that contains the analyte is sprayed through a small-diameter needle. A high positive voltage is applied to this needle to produce a Tayler cone from which droplets of the solution are sputtered. Protons from the acidic conditions give the droplets a positive charge causing them to move from the needle toward the negatively charged instrument. During the course of this movement, evaporation reduces the size of the droplets into a population of smaller, charged droplets. The evaporation and droplet-spotting cycle repeats until the small size and charging of the droplet desorbs protonated molecules into the gas-phase where they can be directed into the mass spectrometer by appropriate electric fields. One characteristic of ESI is that the acidic conditions used to produce the positively charged droplets tend to protonate all available basic sites. In the process, multiprotonated molecules $(M + nH)^{n+}$ that are formed are analyzed giving rise to a series of consecutive peaks at $(M + n)/n$ along the m/z scale. For a single compound, each adjacent pair of peaks must differ by one charge. Thus unknown charge state of the ion can be derived from any two of such pairs and therefore the molecular weight of the biomacromolecule.

For MADI, the biomolecule is dissolved in a solution of a UV (ultraviolet)-absorbing compound (e.g. nicotinic acid or 3,5-dimethoxy-4-hydroxycinnamic acid), referred to as the matrix, and placed on a probe or stage for the mass spectrometer. As the solvent dries, the matrix compound crystallizes and the biomolecule is included in the matrix crystals. Pulses of UV laser light (266 nm for nicotinic acid or 355 nm for cinnamic acid derivative) are used to vaporize small amounts of the matrix and the included biomolecular ions are carried into the gas phase in the process. Ionization occurs by protonation in the acidic environment produced, the acidity of most matrix compounds and by the addition of dilute acid to the sample. The multiprotonation is not difficult to interpret, because their m/z are integral fractions of that for $z = 1$. The molecular weights of biomacromolecules with masses greater than 150 kDa have been analyzed with high accuracy (better than 0.1 %) by the MS technique.

3.4.3 Molecular dimension

The dimensions of a biomacromolecules are also important. If the molecule is rigid and has a regular structure such as a rod or a sphere, dimensions (represented by length and radius respectively) can be assigned. Indeed, a DNA molecule as a rod can be characterized by its length that is proportional to the molecular weight. However, the shape of most biomacromolecules is flexible and variable, thus two parameters, end-to-end distance and radius of gyration, are used to describe average values over the population at a particular instant. The end-to-end distance, h, is the average separation between the two ends of a linear molecule. For a rigid rod it equals the length of the molecule but for a flexible molecule, its value depends upon the molecular weight and the flexibility. For a random coil (i.e. perfectly flexible molecule), h is proportional to $M^{1/2}$. The radius of gyration, R_G is the root-mean-square (rms) average of the distances of all parts of a molecule from its center of mass. That is

$$R_G^2 = \Sigma m_i r_i^2 / \Sigma m_i$$

in which m_i is the mass of the ith element at a distance r_i from the center of mass. R_G can be measured directly in the analysis of the angular dependence of the scattering light from a solution of biomacromolecules (section 3.4.2). For simple geometric shapes and for the random coil, R_G can be calculated.

Geometric shape	R_G	*Measurement*
Sphere:	$(3/5)^{1/2}r$	r = radius of sphere
Rod:	$(1/12)^{1/2}L$	L = length of rod
Random coil:	$(N/6)^{1/2}L$	N = number of units of length L

The value of R_G can be used to estimate the shape of a molecule by comparing the measured value with that expected for a sphere, rod or random coil.

For highly folded proteins whose shape is not a simple geometric figure but nevertheless reasonably well defined, it is often useful to describe a molecule in terms of the sphere or ellipsoid that can contain the molecule. An interesting shape factor known as frictional ratio, f/f_o measures how much the macromolecule deviates from sphericity. f is the frictional coefficient of a macromolecule and f_o is the hypothetical frictional coefficient of the corresponding spherical macromolecule. According to Stokes' law, the frictional coefficient of a macromolecule is related to its viscosity (η) and the effective hydrated radius known as Stokes radius (r) by: $f = 6\pi\eta r$, and Stokes radius can be estimated from gel filtration chromatography according to

$$K_{av} = \alpha - \beta \log r$$

Similarly, the frictional coefficient of a spherical molecule is related to $f_o = 6\pi\eta r_o$, where r_o is the Stokes radius for the corresponding spherical molecule that can be calculated by

$$r_o = [(3\bar{v}M)/(4\pi N_A)]^{1/3}$$

where $\bar{v}$ is partial specific volume (e.g. ~0.72 cm^3/g for proteins), M is molecular weight and N_A is the Avogadro's number ($6.022 \times 10^{23}\,\text{mol}^{-1}$). The ratio,

$$f/f_o = r/[(3\bar{v}M)/(4\pi N_A)]^{1/3}$$

provides a measure of deviation of the macromolecule from the corresponding unhydrated spherical shape. Table 3.14 gives experimental data of some representative biomacromolecules.

3.5 CHARACTERIZATION: SPECIFIC

3.5.1 Melting temperature of DNA

When a solution of duplex DNA is heated above a characteristic temperature, its two complementary strands separate and assume the random coil conformation. This denaturation process can be followed by monitoring an increase in its UV absorbency at 260 nm, known as the hyperchromic effect. This hyperchromic shift that occurs over a narrow temperature range gives rise to a melting curve. The temperature at its midpoint is known as its melting temperature, T_m (i.e. T_m is the temperature at which an increase in half of the maximum absorbency is attained). T_m is dependent on the ionic strength of the solution (the lower the ionic strength, the lower the T_m) and G + C content of DNA. The higher the G + C content of a DNA, the higher its T_m. The melting temperature of a

TABLE 3.14 Dimension of some biomacromolecules

Biomacromolecule	M (Da)	$\bar{v}$ (cm^3/g)	$R_{G,expt}$ (Å)	r (Å)	$R_{G,expt}/R_{G,sphere}$	f/f_o
Lysozyme	13 930	0.70	14.3		1.17	1.21
Myoglobin	16 890	0.74	16.0		1.21	1.05
Serum albumin	66 000	0.73	29.8	33.7	1.42	1.25
Hemoglobin	68 000	0.75		34.0		
tRNA	26 600	0.53	21.7		1.57	
Catalase	225 000	0.73	31.0		1.28	1.15
Myosin	493 000	0.73	468	257	10.4	3.65
DNA	4×10^6	0.55	1170		15.8	
Tobacco mosaic virus	39×10^6	0.73	924		5.28	2.19

Note: 1. $R_{G,sphere}$ is calculated by assuming spherical molecule according to $R_{G,sphere} = (3/5)^{1/2}(3\,Mv/4\pi N_A)^{1/3}$.

2. f/f_o values have been corrected for the hydration, $(1 + \delta)^{1/3}$, where δ is the volume of water of hydration per unit volume of anhydrous macromolecules.

3. From $R_{G,expt}/R_{G,sphere}$ and f/f_o values: Lysozyme, catalase and myoglobin are not strictly spherical but their overall shape is not far from spherical, whereas myosin, DNA and TMV are nonspherical.

4. An idea concerning the flexibility of DNA rod can be gained also by comparing R_G values. For DNA, the mass per unit length is known to be 200 Da/Å. Thus, DNA of 4×10^6 has a length, $L = 2 \times 10^4$ Å and $R_{G,rod} = (1/12)^{1/2}(1 \times 10^4) = 5774$ Å. This gives $R_{G,expt}/\ R_{G,rod}/ = 0.203$ indicating that the DNA molecule is not fully extended but somewhat flexible.

DNA molecule is linearly related to its %G + C (Marmur and Doty, 1962) by $T_m = 69.3 + 0.41$ (%G + C) at 0.2 M Na^+. Knowledge of T_m is essential in various aspects of DNA function and analyses. Predicted T_m for any oligonucleotides can be obtained from www.genseloligos.com/Calculation/calculation.html.

3.5.2 Buoyant density of biomacromolecules

The differences in the buoyant density of biomacromolecules can be exploited for their separation/purification and characterization. The density of DNA is slightly greater than 1.7 g/mL, while the density of RNA is more than 1.8 g/mL. Proteins have densities less than 1.3 g/mL. Therefore, DNA, RNA and proteins can be separated by the CsCl density gradient centrifugation (~200 000 × g), also known as isopycnic centrifugation. Since single-stranded DNA (ssDNA) is denser than double-stranded DNA dsDNA), the isopycnic centrifugation can be used to separate the randomly coiled ssDNA and helical dsDNA, which also forms a narrower band. The buoyant density (ρ) of a DNA molecule, which is related to its G + C content by

$$\rho = 1.660 + 0.098\ (G + C\ \text{mole fraction})$$

can be estimated with the isopycnic centrifugation using CsCl density gradient of 1.6–1.8 g/mL.

3.5.3 Isoelectric pH of proteins

Proteins generally bear numerous ionizable groups, which have different pKs. At a pH characteristic for each protein, known as its isoelectric point (p*I*), the positive charges on the molecule exactly balance its negative charges. At p*I*, the protein carries no net charge and is immobile in an electric field. Therefore IEF offers a convenient means of determining p*I* of a protein molecule, which forms a band at the position corresponding to the pH of the gradient. In solution of moderate salt concentrations, the solubility of a protein as a function of pH is expected to be at a minimum at the protein's p*I* in solution of moderate salt concentrations.

3.5.4 Removal of glycosides from glycoproteins

The determination of the structures of glycosides of glycoproteins are obtained by first removing them from the protein. *O*-glycosides attached to serine and threonine and *S*-glycoside attached to cysteine can be released by an β-elimination using a mild alkaline treatment (Spiro, 1972). The reaction is usually conducted in the presence of sodium borohydride, which reduces the reducing-end of the released oligosaccharide. This reduction prevents alkaline degradation of the oligosaccharide by the peeling reaction, and it determines the carbohydrate residue that is attached to the protein as an alditol. To achieve β-elimination, various strengths of alkaline (0.05–0.5 N NaOH), temperature of 0–5°C and length of 12–216 h are used. The sodium borohydride is 0.15–1.0 M. A standard procedure uses 0.1 N NaOH and 0.3 M $NaBH_4$ at 37°C for 48 h. The β-elimination reaction does not proceed satisfactorily if the glycosylated amino acid is the C- or N-terminal residue. In such cases, the amino or carboxyl group has to be derivatized to eliminate the charge (Spiro, 1972).

The asparagine N-linked glycosides can be cleaved by hydrazinolysis (Takasaki *et al.*, 1982). The glycoprotein is heated at 100°C with anhydrous hydrazine for 8–12 h. The procedure is carried out by suspending 0.2–1 mg of glycoprotein in 0.5–1.0 mL of freshly distilled anhydrous hydrazine. The solution is heated in a sealed tube at 100°C for 8–12 h. The glycoprotein sample usually dissolves after 1 h. Various endoglycosidases such as endo-β-N-acetylglucosaminidase can be used to liberate asparagine N-linked oligosaccharide chains.

3.6 REFERENCES

Aboul-Enein, H.Y. (ed.) (1999) *Analytical and Preparative Separation Methods of Biomacromolecules*, Marcel Dekker, New York.

Aguilar, M.-I. (ed.) (2004) *HPLC of Peptides and Proteins*, Humana Press, Totowa, NJ.

Allen, R.C., Saravis, C.A. and Maurer, H.R. (1984) *Gel Electrophoresis and Isoelectric Focusing of Proteins: Selected Techniques*, de Gruyter, Berlin.

Andrews, A.T. (1986) *Electrophoresis: Theory, Techniques, and Biochemical and Clinical Applications*, 2nd edn, Oxford University Press, New York.

Andrews, P. (1970) *Methods of Biochemical Analysis*, **18**: 2–53.

Barondes, S.H. (1981) *Annual Reviews in Biochemistry*, **50**, 207–31.

Biemann, K. (1992) *Annual Reviews in Biochemistry*, **61**, 977–1010.

Birren, B. and Lai, E. (eds) (1993) *Pulse Field Gel Electrophoresis*, Academic Press, San Diego.

Bowien, B. and Durre, P. (2003) *Nucleic Acids Isolation Methods*, American Scientific Publishers, Stevenson Ranch, CA.

Burckhardt, J. and Birnstiel, M.L. (1978) *Journal of Molecular Biology*, **118**, 61–79.

Deutscher, M.P. (ed.) (1990) *Guide to Protein Purification: Methods in Enzymology*, Vol. 182, Academic Press, San Diego.

Dunbar, B.S. (1987) *Two-dimensional Electrophoresis and Immunological Techniques*, Plenum, New York.

Edward, D.I. (1970) *Chromatography: Principles and Techniques*, Butterworths, London.

Fichmann, J. (1999) *Methods in Molecular Biology*, **112**, 173–174.

Görg, A., Obermaier, C., Boguth, G. *et al.* (2000) *Electrophoresis*, **21**, 1037–53.

Hames, B.D. and Rickwood, D. (eds) (1981) *Gel Electrophoresis of Proteins*, IRL Press, Oxford, UK.

Harris, E.L.V. and Angal, S. (eds) (1989) *Protein Purification Methods: A Practical Approach*, IRL Press, Oxford, UK.

Karas, M. and Hillenkamp, F. (1988) *Analytical Chemistry*, **60**, 2299–3201.

Kirkland, J.J., Truszkowski, F.A. and Ricker, R.D. (2002) *Journal of Chromatography*, **A965**, 25–34.

Kupke, D.W. (1960) *Advances in Protein Chemistry*, **15**, 57–130.

Larsen, B.S. and McEwen, C.N. (eds) (1998) *Mass Spectrometry of Biological Materials*, 2nd edn, Marcel Dekker, New York.

Lis, H. and Sharon, N. (1973) *Annual Reviews in Biochemistry*, **42**, 541.

Marmur, J. and Doty, P (1962) *Journal of Molecular Biology*, **5**, 109–18.

Martin, R.G. and Ames, B.N. (1961) *Journal of Biological Chemistry*, **236**, 1372–79.

McCormick, A., Miles, L.E.M. and Chrambach, A. (1976) *Analytical Biochemistry*, **75**, 314.

McEwen, C.R. (1967) *Analytical Biochemistry*, **20**, 114–49.

Miller, J.M. (2005) *Chromatography: Concepts and Contrasts*, John Wiley & Sons, Hoboken, NJ.

Millner, P. (ed.) (1999) *High Resolution Chromatography: A Practical Approach*, Oxford University Press, Oxford, UK.

Mohr, P. and Pommerening, K. (1985) *Affinity Chromatography: Practical and Theoretical Aspects*, Marcel Dekker, New York.

Nguyen, N.Y. and Chrambach, A. (1980) *Electrophoresis*, **1**, 14.

Rickwood, D. and Hames, B.D. (eds) (1982) *Gel Electrophoresis of Nucleic Acids*, IRL Press, Oxford, UK.

Righetti, P.G. and Bossi, A. (1997) *Analytical Biochemistry*, **247**, 1–10.

Rabilloud, T. (2000) *Analytical Biochemistry*, **72**, 48A–55A.

Rosenberg, I.M. (2005) *Protein Analysis and Purification*, 2nd edn, Birkhaüser, Boston, MA.

Schoemakers, P.J. (1986) *Optimization of Chromatographic Selectivity*, Elsevier, Amsterdam.

Shaw, C. (1998) Solubilization and assay of cellular and tissue protein. *Methods in Molecular Biology*, **105**, 287–93.

Sheehan, D. (2000) *Physical Biochemistry: Principles and Applications*, John Wiley & Sons, New York.

Siegel, L.M. and Monty, K.J. (1966) *Biochimera Biophysica Acta*, **112**, 346–62.

Spiro, R.G. (1972) *Methods in Enzymology*, **28**, 35–40.

Strege, M.A. and Lagu, A.L. (2004) *Capillary Electrophoresis of Proteins and Peptides*, Humana Press, Totowa, NJ.

Takasaki, S., Mizuochi, T. and Kobata, A. (1982) *Methods in Enyzmology*, **83**, 263–8.

Turkova, J. (ed.) (1978) *Affinity Chromatography*, Elsevier, New York.

Ui, N. (1979) *Analytical Biochemistry*, **97**, 65–71.

Watson J.T. (1997) *Introduction to Mass Spectrometry*, 3rd edn, Lippinocott-Raven, PA.

Weber, K., Pringle, J.R. and Osborn, M. (1972) *Methods in Enzymology*, **26**, 3–27.

Westermeier, R. (2001) *Electrophoresis in Practice: A Guide to Methods and Applications of DNA and Protein Separations*, 3rd edn, Wiley-VCH, New York.

Whitehourse, C.M., Dreyer, R.N., Yamashita, M. *et al.* (1985) *Analytical Chemistry*, **57**, 675–9.

World Wide Webs cited

BioProtocol:	http://www.bio.com/protocolstools/protocol.jhtml.
ExPaSy protocol:	http://expasy.cbr.nrc.ca/ch2d/protocols/
LabVelocity:	http://reserchlink.labvelocity.com/tools/index.jhtml
Oligonucleotide calculation:	http://www.genseloligos.com/Calculation/calculation.html

CHAPTER 4

BIOMACROMOLECULAR STRUCTURE: NUCLEIC ACIDS

4.1 STRUCTURAL ORGANIZATION

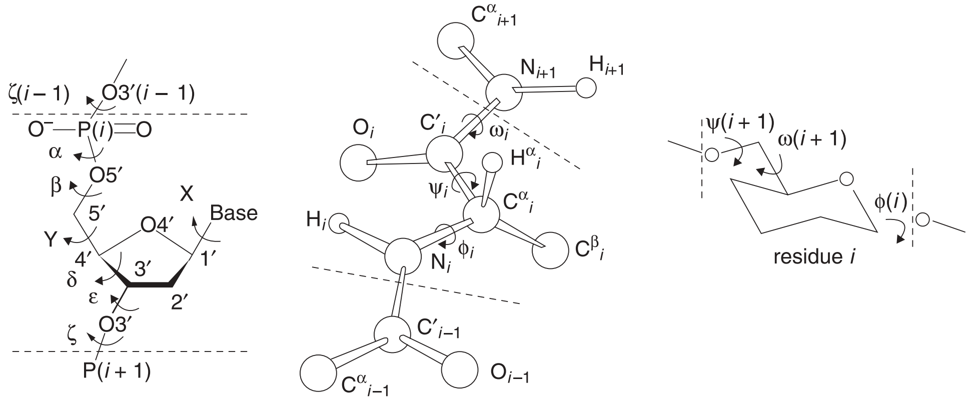
4.1.1 Structural hierarchy

The structure of biological macromolecules is hierarchical, with distinct levels of structural organization representing increased levels of complexity and are defined as follows:

- *Monomers* are the simple building blocks that when polymerized, yield a macromolecule. These include nucleotides, amino acids and monosaccharides of the biomacromolecules (Figure 4.1).
- *Primary structure* (1° structure) is the arrangement (or sequence) of monomeric residues in the covalently linked biopolymers. This arrangement is linear for nucleic acids and proteins, but can be branched in polysaccharides.
- *Secondary structure* (2° structure) is the local regular structure of a macromolecule or specific region of the molecule. These are the helical, pleated and coil structures. Thus secondary structure includes all three-dimensional regions that have ordered locally symmetric backbone structure.
- *Tertiary structure* (3° structure) describes the global three-dimensional fold or topology of the macromolecule, relating the positions of each atom and residue in three-dimensional space. For biomacromolecules with a single subunit, the functional tertiary structure is its native structure. Thus the tertiary structure includes a description not only of local symmetric structure (2° structure) but also of the spatial location of all residues as possible.
- *Quaternary structure* (4° structure) is the spatial arrangement of multiple distinct polymeric chains (subunits) that form a functional complex.
- *Quinternary structure* (5° structure) refers to the association of one class of biomacromolecule with another class of biomacromolecule to form complexes of cellular components such as histone (DNA-protein), ribosome (RNA-protein) and glycoprotein (oligosaccharide-protein).

Not all levels of structure are required or represented in all biomacromolecules. In general, all biomacromolecules require a level of structure up to and including secondary/tertiary for biological function. It should be emphasized that a representation/visualization of a molecule is only a model described by the atoms and the positions (coordinates) of the atoms in three-dimensional space. This model is correct only when it conforms to the experimentally observed properties and activities.

Biomacromolecules, by C. Stan Tsai
Copyright © 2007 John Wiley & Sons, Inc.



A. Nucleotide unit B. Peptide unit C. Hexopyranose unit

Figure 4.1 Covalent linkages of monomer units in biomacromolecules. Covalent linkages of monomer units showing torsion angles (ϕ and ψ) which affect the main chain conformations of biopolymers are illustrated for polynucleotide (A), polypeptide (B) and polysaccharide (C) chains according to the IUBMB notation. The two torsion angles, ϕ and ψ, specified around the phosphodiesteric bonds of nucleic acids correspond to α and ξ, respectively

4.1.2 Representation of structures of nucleic acids

Nucleic acids, which are polymers of deoxyribonucleotides/ribonucleotides (Blackburn and Gait, 1997; Bloomfield *et al.*, 2000; Neidle, 2002), are abbreviated as

d· · · · · ·$_{p}N_{p}N_{p}N\ {}_{p}N_{p}N_{p}N_{p}$· · · · · · for DNA where N = A, G, C and T

and r · · · · · ·$_{p}N_{p}N_{p}N\ {}_{p}N_{p}N_{p}N_{p}$· · · · · · for RNA where N = A, G, C and U

The prefixes d and r, which represent deoxyribose of DNA and ribose of RNA respectively, are omitted where an implication of the type of polynucleotides is obvious from the context of nucleotides involved (i.e. T for DNA, while U for RNA). In the linear code representation, the nucleotide sequence is written from the left for the 5′-end to the right for the 3′-end (e.g. coding nucleotide sequence for human lysozyme (Fasta format))

ATGGCCCTGTGGATGCGCCTCCTGCCCCTGCTGGCGCTGCTGGCC
CTCTGGGGACCTGAC

CCAGCCGCAGCCTTTGTGAACCAACACCTGTGCGGCTCACACCTGGTG
GAAGCTCTCTAC

CTAGTGTGCGGGGAACGAGGCTTCTTCTACACACCCAAGACCCGC
CGGGAGGCAGAGGAC

CTGCAGGTGGGGCAGGTGGAGCTGGGCGGGGGCCCTGGT
GCAGGCAGCCTGCAGCCCTTG

GCCCTGGAGGGGTCCCTGCAGAAGCGTGGCATTGTGGAACAATGCTG
TACCAGCATCTGC

TCCCTCTACCAGCTGGAGAACTACTGCAACTAG

Nucleotide sequences of nucleic acids can be retrieved from GenBank of the National Center for Biotechnology Information (NCBI) of the National Institute of Health (NIH) at http://www.ncbi.nlm.nih.gov/Genbank, European Bioinformatics Institute (EBI) of the European Molecular Biology Laboratory (EMBL) at http://www.ebi.ac.uk and DNA Data Bank of Japan (DDBJ) at http://www.ddbj.nig.ac.jp.

Almost all RNAs consist of a single strand. An important exception is the RNA of some tumor viruses, which seems to exist as a dimer of two identical strands. The most common DNAs contain two strands of the same length, with complementary sequences allowing for the full A-T and G-C pairings along the entire length. Thus only the nucleotide sequence of one strand (5′→3′) of the DNA duplex is normally given, since the sequence of the complementary strand (3′→5′) is understood/implicated according to the base pairings. Some natural DNA strands are linear and some are circular, while a few DNA duplexes can be shown to contain single-strand breaks. The simplified sketches of these DNA chain structures that often appear in some texts are depicted in Table 4.1.

4.2 SEQUENCE ANALYSIS OF NUCLEIC ACIDS

4.2.1 General

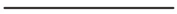
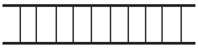
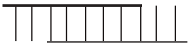
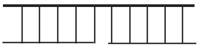
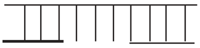
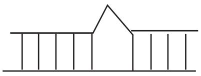
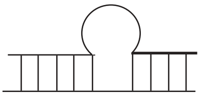
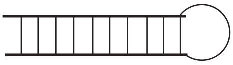
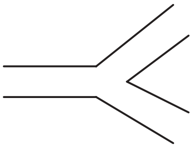
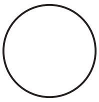
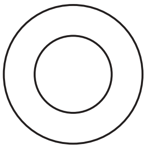
Nucleic acids, both DNA and RNA, have each nucleotide joined by a phosphodiesteric linkage from its 5′-hydroxyl group to the 3′-hydroxy group of next nucleotide in their regular primary structures. This means that the uniqueness of a given nucleic acid primary structure resides solely in the sequence of bases. The basic strategy of nucleic acid sequencing (Ansorge *et al*., 1997; Brown, 1994) is identical to that for other biomacromolecules, namely:

- the specific and reproducible cleavage of the biopolymer chain into fragments and fractionation of the fragments of manageable size to be fully sequenced;
- the sequencing of the individual fragments;
- the repetition of the preceding steps using a degradation procedure that yields a set of the fragments that overlap the cleavage sites in the previous set(s);
- the ordering of the fully sequenced fragments using the information from the overlapped cleavage sites via different degradation procedures.

The development of DNA sequencing has been greatly advanced by two breakthroughs:

1. The discovery of restriction endonucleases (restriction enzymes) that cleave DNA at specific oligonucleotide sites, generating unique fragments of manageable size. The type II restriction enzymes (Table 4.2) that cleave the DNA within the recognition sequences are most widely used for cutting DNA chain at specific sequences (identical to the recognition sequences).
2. The high-resolution polyacrylamide gel electrophoresis (hrPAGE) that resolve DNA fragments that differ from one another in length by one nucleotide, leading to the development of the ladder sequencing technique. Virtually all current DNA sequencing methods are based on the ability to fractionate single-stranded DNA by gel electrophoresis in the presence of a denaturant with single base resolution. Information about the location of particular bases in the sequence is converted into a specific DNA fragment size. Then these fragments are separated and analyzed by hrPAGE. The denaturant, usually 7M urea, is included to eliminate most of the secondary structure that individual DNA strands can achieve by intramolecular base pairing. To a good approximation, the mobility, v of oligonucleotides in denaturing PAGE

TABLE 4.1 Sketches of some DNA chain structures

Chain structure	Sketch
Linear, single strand	
Linear, double strand with blunt ends	
Linear, double strand with cohesive ends	
Linear, nicked double strand	
Linear, gapped double strand	
Double strand, bulge	
Double strand, loop	
Hairpin loop	
Branched, e.g. Y-branch	
Circular, single strand	
Circular, doulbe strand	

is proportional to 1/L(n), where L(n) is the length of the molecule (number of nucleotides).

Two basic protocols for DNA sequencing are commonly adapted in laboratories; the chemical cleavage method (Maxam and Gilbert, 1977) and the enzymatic chain termination or dideoxy method (Sanger *et al.*, 1977). Of the two approaches, the Sanger dideoxy

TABLE 4.2 Selected example of type II restriction endonucleases and their cleavage sequences. The recognition sequences of type II restriction endonucleases are shown with the arrow (↓) indicating the cleavage site

Name	Sequence	Name	Sequence
Acc I	GT↓MKAC	Mlu I	A↓CGCGT
Alu I	AG↓CT	Msp I	C↓CGG
Apy I	CC↓WGG	Nco I	C↓CATGG
Bal I	TGG↓CCA	Nde I	CA↓TATG
BamH I	G↓GATCC	Nru I	TCG↓CGA
Ban III	AT↓CGAT	Pst I	CTGCA↓G
Bgl I	GCCN$_4$↓NGGC	Pvu II	CAG↓CTG
Bgl II	A↓GATCA	Rsa I	GT↓AC
Cfo I	GCG↓C	Sal I	G↓TCGAC
Cla I	AT↓CGAT	Sca I	AGT↓ACT
Dra I	TTT↓AAA	Sph I	GCATG↓C
EcoR I	G↓AATTC	Sse I	CCTGCA↓GG
EcoR II	↓CCWGG	Ssp I	AAT↓ATT
Fsp I	TGC↓GCA	Sun I	C↓GTACG
Hae III	GG↓CC	Taq I	T↓CGA
Hha I	GC↓GC	Tha I	CG↓CG
Hinc II	GTY↓RAC	Vsp I	AT↓TAAT
Hind III	A↓AGCTT	Xba I	T↓CTAGA
Kpn2 II	T↓CCGGA	Xho I	C↓TCGAG

Notes: 1. Restriction endonucleases are named by italicized three-letter codes of which the first capital letter denoting the genus of the organism of origin and the next two letters are an abbreviation of the particular species. The following capital letter and number codes represent the strain and sequence as the enzymes are identified.

2. R for A or G; M for C or A; K for T(U) or G; W for T(U) or A; N for any base.

3. If we assume equimolar proportions and random distribution for the four nucleotides in DNA, a particular sequence should occur every 256 (4^4) for the tetranucleotide or 4096 (4^6) for the hexanucleotide. Therefore the fragments generated by the four-cutter and six-cutter enzymes should average about 250 bp and 4000 bp in length respectively.

method is now the most widely used technique for sequencing DNA. Several texts have reviewed many variations made to this sequencing technique (Alphey, 1997; Ansorge *et al*., 1997), but the principle described below remains the same.

4.2.2 Chemical cleavage method (Maxam and Gilbert, 1980)

In the Maxam–Gilbert sequencing method, the ends of the DNA are distinguished by specifically labeling one of them. Usually this is done directly and covalently with a kinase that places a radiolabeled phosphate (^{32}P) at the 5′-terminus of the template. The base-specific or base-selective partial chemical cleavage is used to cleave the end-labeled DNA at only one type of nucleotide under conditions where there is an average of only one cut per template molecule with each cleavage scheme employed. This is carried out under conditions to produce a very broad range of fragment sizes whose members extend from the ^{32}P-labeled end to one of the positions occupied by the cleaved base that reflects the entire sequence of the template. Four separate chemical fragmentation reactions are carried out (Table 4.3); each one cleaves at specific base(s).

TABLE 4.3 Protocol for chemical cleavage of DNA chain

Specific site(s) of cleavage	Reactions/reagents used	Ladder
A + G	Depurination by protonation of purines with HCOOH followed by piperidine cleavage of the chain at G and A	A + G
G	Depurination of N_7-methylG with dimethyl sulfate followed by piperidine cleavage of the chain at G	G
C + T	Nucleophilic split of pyrimidine ring with hydrazine followed by piperidine cleavage of the chain at C and T	C + T
C	Suppression of hydrazine attack on T with NaCl (2 M) followed by piperidine cleavage of the chain at C	C

The fragments are fractionated and the sizes of the labeled pieces are measured, usually in four parallel electrophoretic lanes. The pattern of bands seen is often called a ladder. To read base sequence form 5′-end beginning with the bottom-most band, proceed upward remembering that the two ladders, A + G and C + T contain all bands that arise from partial cleavage of the 5′-end labeled DNA. If a band appears in the A + G ladder, it derives from cleavage at a purine. This purine is a G if it falls under G ladder, otherwise it is an A. Likewise the C + T ladder indicates a pyrimidine, and if it is also under C ladder, that pyrimidine is a C, if not it is a T.

4.2.3 Enzymatic chain termination/dideoxy method (Smith, 1980)

The second general approach to DNA fragmentation for ladder sequencing developed by Sanger and coworkers is in widespread use today for its ease of adapting to four-color fluorescent detection. It starts with a single-stranded template. A primer is annealed to this template, near the 3′-end of the DNA to be sequenced. The primer must correspond to a known DNA sequence, either in the target or more commonly, in the flanking vector sequence. A DNA polymerase is used to extend the primer in a sequence-specific manner along the template. However, the sequence extension is halted, in a base specific manner, by allowing the occasional uptake of chain terminators: deoxynucleoside triphosphate (dpppN, dNTP) analogs that cannot be further extended by the enzyme. Almost all current DNA sequencing uses dideoxynucleoside triphosphates (dd-pppNs, ddNTPs) as terminators. These derivatives lack the 3′ OH needed to form the next phosphodiester bond. Four separate chain extension reactions are carried out, each one with a different terminator. A label can be introduced in several different ways, such as through the primer, the ddNTP terminator or internal dNTPs. The resulting mixture of DNA fragments is melted off the template and analyzed by gel electrophoresis.

Specifically a DNA polymerase extends an oligonucleotide primer annealed to a unique location on a DNA template by incorporating deoxynucleotides (dNTPs) complementary to the template. Synthesis of this new DNA strand continues until the reaction is randomly terminated by the inclusion of a dideoxynucleotide (ddNTP). This results in a population of truncated sequencing fragments of varying length. The identity of the chain-terminating nucleotide at each position is specified by running four separate base-specific reactions, each of which contains a different dideoxynucleotide. The four such fragment sets are loaded in adjacent lanes of a polyacrylamide gel and separated according to the fragment size by electrophoresis, which resolves DNA fragments differing in length, by just one nucleotide. If a radioactive label is introduced into the sequencing reaction products, then autoradiographic imaging of the DNA band pattern in the gel can be used to deduce the DNA sequence. The smallest fragments migrate fastest and because fragments differing by only single nucleotides in length are readily resolved by hrPAGE. The electrophoregram of the gel can be read from bottom to top in the 5′ → 3′ direction.

4.2.4 Mass spectrometric analysis

For the sequence analysis of RNA, the sample is first transcribed by a reverse transcriptase to cDNA, which is then analyzed by either the chemical cleavage or enzymatic termination methods. These techniques are limited by the premature termination arising from RNA secondary structure and inapplicability to analyzing modified bases. Mass spectrometry (Larsen and McEwen, 1998) offers an alternative technique applicable to modified bases. The molecular weight determination by electrospray ionization (ESI) or matrix (e.g. 3-hydroxypiconic acid or 2,4,6-trihydroxyacetophenone)-assisted laser desorption ionization (MALDI) mass spectrometry (subsections 3.4.2 and 5.3.1) may furnish preliminary information about the likely structure of an oligonucleotide. However, tandem mass spectrometry (subsection 5.3.2) provides a powerful tool for analyzing the structure and sequence of RNAs (Thomas and Akoulitchev, 2006). Tandem mass spectrometry (MS–MS) enables a single oligonucleotide ion produced from either ESI or MALDI to be selected for fragmentation by collisionally activated dissociation (CAD) or collisionally induced dissociation (CID). This produces a unique series of fragment ions (m/z), which are then analyzed.

The proposed nomenclature (Wu and McLuckey, 2004) for describing the fragmentation and m/z ions for the nucleotide cleavages is given in Figure 4.2. Four cleavage sites along the phosphodiesteric backbone of the oligonucleotide chain are designated a, b, c and d if the fragment ion contains the 5′ terminus, or w, x, y and z if the fragment ion contains the 3′ terminus. The subscript gives the number of nucleotides from the respective termini.

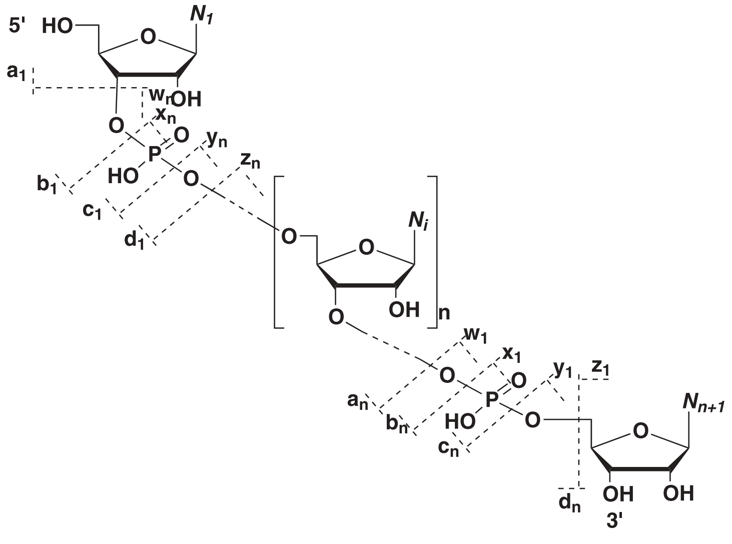


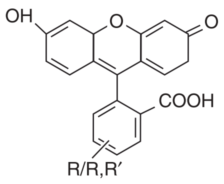
Figure 4.2 MS fragmentation of oligonucleotides. Four possible fragmentation sites of oligonucleotides and their proposed nomenclature are depicted

4.2.5 Automated DNA sequencing technology

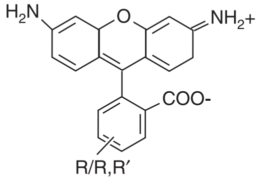
The efficient sequencing of DNA is now a reality due to the development of fluorescence-based dideoxynucleotide sequencing chemistries coupled with instrumentation for real time detection of dye-labeled DNA fragments during gel electrophoresis. In the automated DNA sequencing system (Adams *et al*., 1994), the reaction products are labeled with an appropriate fluorescent dye (Smith *et al*., 1986; Brumbaugh *et al*., 1988), and DNA fragments are detected upon irradiation with a laser as they move through the electrophoresis gel. A detector collects the fluorescence emission and the resultant signal produces a trace pattern, which correlates to a DNA sequence. In automated fluorescent detection, the sample is examined at a constant distance from the starting point, D. The time it takes a fragment of a particular length (L) to reach this distance is proportional to $D/v = DL$ where the velocity, v of DNA fragments in denaturing acrylamide gel electrophoresis is proportional to $1/L$. Hence the spacing between two bands of length L and $L - 1$ is $DL - D(L - 1) = D$, i.e. independent of size, but it can be increased by using longer running gels.

For DNA sequencing, four different colored fluorescent dyes are commonly used as detection schemes. One dye is used for each base-specific primer extension. The ideal set of dyes would have similar chemical structures so that their presence would affect the electrophoretic mobility of labeled DNA fragments in identical ways. They would also have emission spectra as distinct as possible and they would all be excitable by the same wavelength so that a single excitation source would suffice for all four dyes. The dyes would also allow similar high sensitivity detection so that signal intensities from the four different cleavage reactions would be comparable. All currently used dyes for four-color DNA sequencing are exited in the UV/visible range.

Commercialized automated DNA sequencing instruments employ either four distinct dye-labeled primers with non-fluorescent terminators per DNA sample or one non-fluorescent primer with four distinct fluorescent terminators per DNA sample. The first commercialized near IR dyes introduced for automated DNA sequencing were IRDye41 and IRDye40. These dyes are from the heptamethine carbocyanine dye family and nominally absorb and fluoresce near 800 nm. The isothiocyanate (NCS) functionality is used to couple the dye via thiourea linkage to an amino linker located at the 5′ end of the primer. A phosphoramidite derivative (IRDye800) provides direct labeling of DNA primers using an automated DNA synthesizer. For dye labeled terminator chemistry, the IRDye40 is attached to bases linked to a triphosphate through an acyclo bridge. The incorporation of this substrate terminates DNA chain elongation. Two fluorescein derivatives (FAM and JOE) and two rhodamine dye derivatives (TAMRA and ROX) are now in use for four visible dye primer-based DNA sequencing.



fluorescein dyes



rodamine dyes

The four visible color dye terminator-based DNA sequencing uses rhodamine dye derivatives, R110, R6G, TAMRA and ROX or their 4,7-dichloro-substituted derivatives (dR110, dR6G, dTAMRA and dROX). Their spectral properties of these dyes are listed in the Table 4.4.

TABLE 4.4 Spectral properties (aqueous solution) of several commercial dyes for DNA sequencing

Dye	Absorption max. (nm)	Emission max. (nm)
FAM	490–495	510–520
R110	500–505	525–530
dR110		530–535
JOE	520–525	550–555
R6G	525–530	555–560
dR6G		560–565
TAMRA	550–555	580–585
dTAMRA		590–595
ROX	580–585	605–610
dROX		615–620
Cy5	650–655	665–670
IRDye700	685–690	710–715
IRDye40	765–770	785–790
IRDye41	795–800	820–825
IRDye800	795–800	820–825

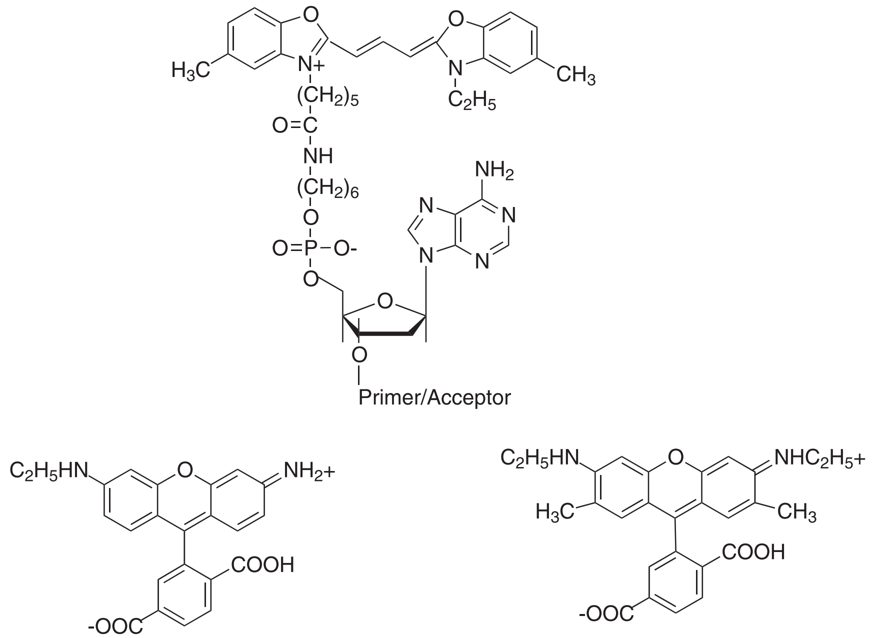
A significant improvement in fluorescent dyes for automated sequencing is the use of energy transfer method (Glazer and Mathies, 1997). Primers contain a pair of fluorescence dyes (Figure 4.3). One dye is common to all four primers. This is optimized to absorb the exciting laser dyes. The second dye is different in each primer and is close enough in each case that fluorescence resonance energy transfer is 100% efficient. Thus all the excitation energy migrates to the second dye where it is subsequently emitted. The second dyes are chosen so that they have as different emission spectra as possible to maximize the ability to accurately discriminate the four different colors.

4.3 SECONDARY STRUCTURE AND STRUCTURE POLYMORPHISM OF DNA

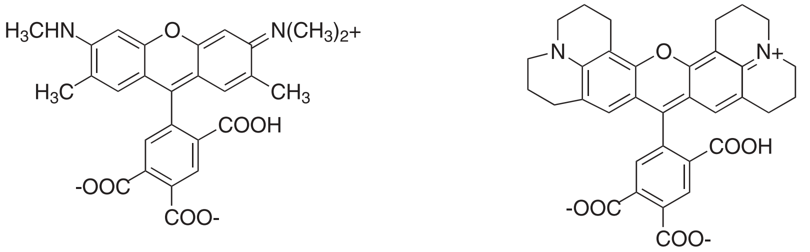
4.3.1 Key structural features of nucleic acids

The critical feature of DNA is its linear order of the four nucleotides. The structure of DNA physically protects the all-important atoms of the bases from chemical modification by environment. In cells, DNA occurs predominantly in antiparallel double stranded forms; plus-strand in $5' \rightarrow 3'$ and minus strand in $3' \rightarrow 5'$ directions. The two strands are helically coiled, maximizing the exposure of the charged and polar backbone to water, while the aromatic bases lying within the middle are shielded. The complementary arrangement of hydrogen bond donors and acceptors allow the Watson–Crick base-pairing of adenine (A) with thymine (T), and guanine (G) with cytosine (C), known as canonical pairs. This minimizes the strain in the backbone of the DNA double helix and maximizes the number of hydrogen bonds possible in the two base pairs (bp). Various canonical and non-canonical base pairings can be viewed at http://prion.bchs.uh.edu/bp_type/bp_structure.html.

The sugars and phosphate are located toward one side of the A—T and G—C base pairs. In a typical DNA duplex, the phosphodiester backbones are on the outside of the structure, while the bps are internal. The spaces between the two backbones are called


H3C
N+
(CH2)5
O=C
NH
(CH2)6
O
O=P-O-
O
O
CH3
N
C2H5
NH2
N
N
N
N
O
Primer/Acceptor
C2H5HN
NH2+
COOH
-OOC
C2H5HN
NHC2H5+
H3C
CH3
COOH
-OOC

(a) R110 (b) R6G


H3CHN
N(CH3)2+
H3C
CH3
COOH
-OOC
COO-
N
O
N+
COOH
-OOC
COO-

(c) TAMRA (d) ROX

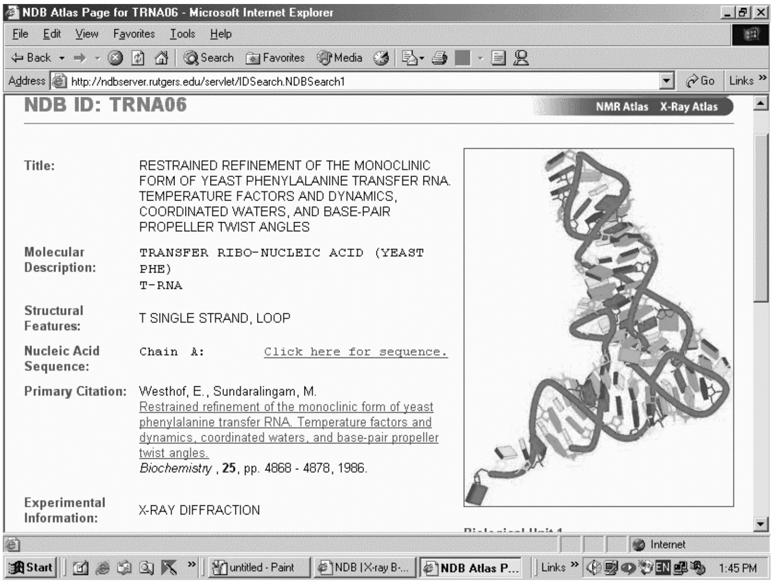
Figure 4.3 Structures of the donor (CYA) and four different acceptor fluorescent dyes (a – d) that can be detected simultaneously in DNA sequencing

grooves. Usually one groove is much broader (the major groove) than the other (the minor groove). The edge of a bp from the sugar C1′ along the N3 edge of a purine (Pu), and C2 edge of a pyrimidine (Py) to the sugar C1′ is located in the minor groove. The other edge of the bp from the sugar C1′ along the C6 of a Pu and C4 edge of Py to its sugar C1′ is located in the major groove. More than one type of helix can be formed by DNA and the major groove and minor groove vary in depth and width, depending on the type of helix formed. Significantly, the arrangement of hydrogen bond donors and acceptors in the major groove and minor groove provides a means for the recognition of DNA sequences by other molecules, especially DNA-binding proteins.

There are two preferred orientation (*anti* and *syn*) of the base with respect to the sugar, as determined by steric restrictions during the torsion (χ) about the C1′—N glycosyl bond. The sugar is nonplanar, and two of its common puckering modes are C2′-endo (*S*-type) and C3′-endo (*N*-type). The secondary structure of DNA can be described by a number of parameters that define the helix (Dickerson *et al*., 1989):

- *Helix sense* refers to the helical rotation of the double helix. The structure described by Watson and Crick (Watson and Crick, 1953) is a right-handed (clockwise) helix. Most helical forms of DNA are right-handed except Z-DNA, which is left-handed.
- *Residues per turn* refers to the number of bps in one helical turn of DNA, that is, the number of bases needed to complete one 360° rotation. The B-form DNA described by Watson and Crick contains 10 bp per turn. DNA in solution contains 10.4–10.5 bp per turn, although this value can vary considerably as a function of base composition.
- *Axial rise* is the distance between adjacent planar bases in the DNA double helix. In B-form DNA there are about 3.4 Å between adjacent base pairs.
- *Helix pitch* is the length of one complete helical turn of DNA. In B-form DNA, one helical turn of 10 bp is completed in 34 Å.
- *Base pair tilt* refers to the angle of the planar bases with respect to the helical axis. The tilt angle is measured by considering the angle made by a line drawn through the two hydrogen bonded base relative to a line drawn perpendicular to the helix axis. A bp that was perfectly flat, that is, perpendicular to the helix axis, would have a tilt angle of 0°. In B-form DNA, the bps are tilted by only –6°. In A-form DNA, the bps are significantly tilted at an angle of 20°.
- *Base pair roll* refers to the angle of deflection of a bp with respect to the helix axis along a line drawn between two adjacent bps relative to a line drawn perpendicular to the helix axis.
- *Propeller twist* refers to the angle between the planes of two bps. A bp is rarely a perfect flat plane with each aromatic base in the same plane. Rather, each base has a slightly different roll angle with respect to the other base.
- *Diameter of the helix* refers to the width in Å across the helix. B-DNA has a diameter of 20 Å.
- *Rotation per residue*, or *twist angle*, designed as h, refers to the angle between two adjacent base pairs. B-form DNA of Watson–Crick with 10 bp in one 360° helix turn of DNA; the rotation per residue is $h = 36°$. For B-form DNA in solution with 10.5 bp per turn, $h = 34.3°$.

Almost all known DNA and RNA structures can be viewed and retrieved from the Nucleic Acid Database (NDB) at http://ndbserver.rutgers.edu (Figure 4.4). The structure of DNA helix is not at all uniform and monotonous. Moreover, the DNA is a dynamic molecule that can undergo a wide variety of rearrangements in its secondary structure. There are local, sequence-dependent modulations of structures, which are primarily associated with changes in the orientation of bases. Such changes seek to minimize nonbonded interactions between bases and maximize base stacking. They are generally tolerated by the relatively flexible sugar-phosphate backbone. Regular DNA structures are described by a range of characteristic features. The global parameters of average rise D_z and helix rotation Ω per bp define the pitch of the helix. Sideways tilting of the base pairs through tilt angle τ permits the separation of the bases along the helix axis D_z, to be smaller than the van der Waals distance of 3.4 Å, and so gives a shorter, fatter cylindrical envelope for DNA. The angle τ is positive for A-DNA (positive means a clockwise rotation of the bp when viewed end-on and toward the helix axis) but is smaller and negative for B-DNA helices. At the same time, the bps are displaced laterally from the helix axis by a distance


NDB Atlas Page for TRNA06 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://ndbserver.rutgers.edu/servlet/IDSearch.NDBSearch1
Go Links
NMR Atlas X-Ray Atlas
Internet
Start
untitled - Paint
NDB | X-ray B-...
NDB Atlas P...
Links
EN
1:45 PM

NDB ID: TRNA06

Title:	RESTRAINED REFINEMENT OF THE MONOCLINIC FORM OF YEAST PHENYLALANINE TRANSFER RNA. TEMPERATURE FACTORS AND DYNAMICS, COORDINATED WATERS, AND BASE-PAIR PROPELLER TWIST ANGLES
Molecular Description:	TRANSFER RIBO-NUCLEIC ACID (YEAST PHE) T-RNA
Structural Features:	T SINGLE STRAND, LOOP
Nucleic Acid Sequence:	Chain A: Click here for sequence.
Primary Citation:	Westhof, E., Sundaralingam, M. Restrained refinement of the monoclinic form of yeast phenylalanine transfer RNA. Temperature factors and dynamics, coordinated waters, and base-pair propeller twist angles. *Biochemistry* , **25**, pp. 4868 - 4878, 1986.
Experimental Information:	X-RAY DIFFRACTION

Figure 4.4 Retrieval of nucleic acid structure from Nucleic Acid Database. Nucleic acid structures in pdb format can be retrieved from NDB (http://ndbserver.rutgers.edu/index.html) by entering NDB id or PDB id followed by clicking the Biological unit coordinate. Alternatively the DNA and RNA structures can be selected from Atlas as exemplified for tRNA^{phe} (TRNA06 or 1TRA)

D_a. This parameter, together with the groove width, defines the depth of the major and minor grooves.

4.3.2 DNA polymorphism

In addition to differences at the bp level, the overall helical structure (secondary structure) of DNA is also polymorphic, adopting several distinct conformations (Figure 4.5). Some characteristics of polymorphic DNA helical structures (Table 4.5) are summarized below.

4.3.2.1 B-form DNA. The predominant DNA structure found under physiological conditions is referred to as the B-form, also known as Watson and Crick double-stranded DNA structure. Two important pieces of information were critical for the development of the Watson and Crick structure (Watson and Crick, 1953). ‘Chargaff’s rules’ stated that the amount of adenine always equaled the amount of thymine, and the amount of guanine always equaled to the amount of cytosine (Chargaff, 1951). The second piece of information came from X-ray diffraction patterns of DNA fibers (sodium salt of DNA fibers at 92% relative humidity), which showed that the geometric shape of DNA is a right-handed helix (Wilkins *et al*., 1953). These observations lead to the proposed double-stranded DNA structure, which contains two antiparallel strands of polydeoxynucleotide chain connected by Watson–Crick A—T and G—C bps (canonical pairs) that spiral around a central polymer axis (Figure 4.5B). The specific nature of Watson–Crick base pairing results in a duplex composed of single-strands that are self-complementary; thus knowledge of the nucleotide sequence in one strand is sufficient to define the primary sequence of the other, a feature that facilitates the replication and repair of DNA. B-Form DNA adopts a right-handed helical structure containing a hydrophobic interior of Watson–Crick bps stacked

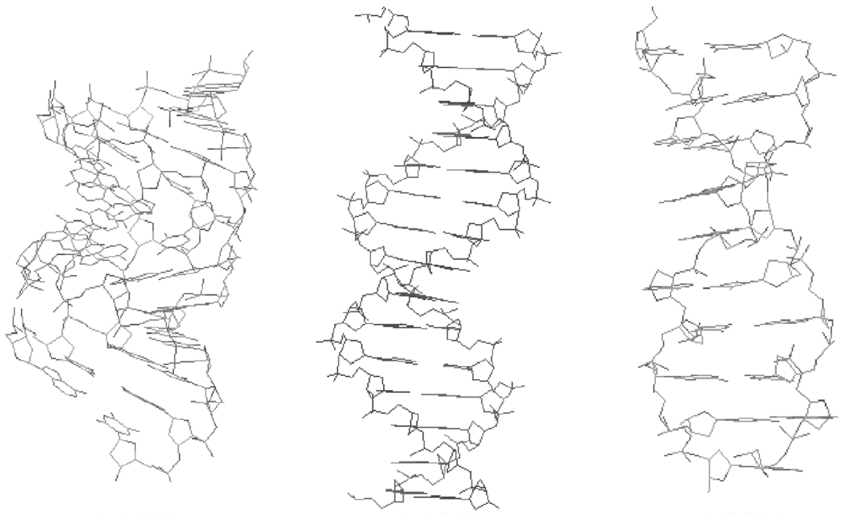

A: A-DNA B: B-DNA C: Z-DNA

Figure 4.5 Major helical structures of DNA duplex. Three major DNA double-stranded structures; A (A-form), B (B-form) and C (Z-form), showing their main structural features are constructed with HyperChem as illustrated

TABLE 4.5 Helical parameters for polymorphic DNA secondary structures

Struct Type	Resid/ turn	Twist (t/°)	Displace (D/Å)	Rise (Å)	Tilt (τ°)	Groove width (Å)		Groove depth (Å)	
						minor	major	minor	Major
A-DNA	11	32.7	4.5	2.56	20	11.0	2.7	2.8	13.5
B-DNA	10	36	–0.2/–1.8	3.4	–6	5.7	11.7	7.5	8.8
C-DNA	9.33	38.5	–1.0	3.31	–8	4.8	10.5	7.9	7.5
D-DNA	8	45	–1.8	3.03	–16	1.3	8.9	6.7	5.8
T-DNA	8	45	–1.43	3.4	–6				
Z-DNA	12	–9,–51	–2/–3	3.7	–7	2.0	8.8	13.8	3.7
A-RNA	11	32.7	4.4	2.8	16/19				
A′-RNA	12	30	4.4	3.0	10				

Notes: 1. All are right-handed helices except Z-DNA, which is left-handed helix.

2. Sugar pucker for most of them are C3′-endo except B-DNA (C2′-endo), C-DNA (C3′-exo) and qT-DNA (C2′-exo).

3. Three major forms of DNA helical structures (secondary structures) have been found and well-studied are B, A, and Z forms:

Parameter	A-DNA	B-DNA	Z-DNA
Helix sense	Right	Right	Left
Residue per turn	11	10 (10.5)	12
Axial rise (Å)	2.55	3.4	3.7
Helix pitch (°)	28	34	45
Base pair tilt (°)	20	–6	7
Rotation per residue (°)	33	36 (34.3)	–30
Diameter of helix (Å)	23	20	18
Glycosidic bond configuration			
dA, dT, dC	anti	anti	anti
dG	anti	anti	syn
Sugar pucker			
dA, dT, dC	C3′ endo	C2′ endo	C2′ endo
dG	C3′ endo	C2′ endo	C3′ endo
Intrastrand phosphate-phosphate distance (Å)			
dA, dT, dC	5.9	7.0	7.0
dG	5.9	7.0	5.9

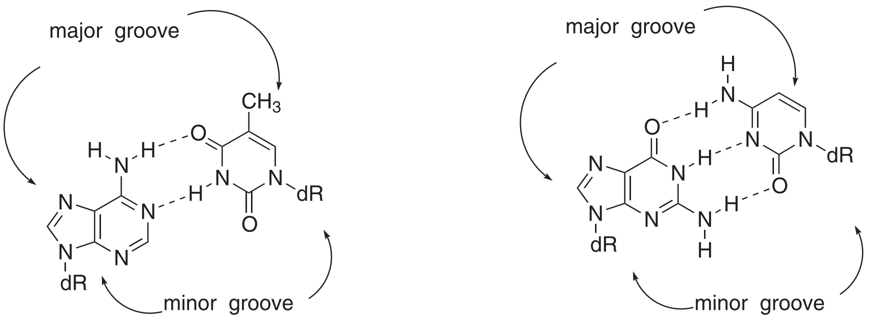


Figure 4.6 Watson–Crick pair pairing and grooves

nearly perpendicular to the central axis at 0.34 nm intervals. Each bp plane of B-form DNA is rotated approximately 36° relative to the one preceding it, resulting in a complete right-handed helical turn for every 10 contiguous base pairs (10.5 bp per turn in solution) and thus a helical pitch of ~3.4 nm. With the Watson–Crick bps inside, the anionic sugar-phosphate backbone spirals around the outside of the helix, creating a hydrophilic exterior with a net charge of −2 for each repeat unit. The deoxyribose ring adopts a C2′-endo (*S*-type) conformation, which the *N*-glycosidic bond angle is in an *anti*-configuration.

The overall structure of B-form DNA creates two distinct helical grooves, the minor and the major (Figure 4.6), which spiral around the surface of the double strand. In B-DNA, the minor groove is narrow, while the major groove is wide, with both grooves possessing a moderate, nearly equivalent depth. Importantly these two grooves create unique microenvironments for the binding and recognition of ligands (i.e. proteins or small molecules).

The floor of the major and minor grooves is defined by the opposite sides of the stacked Watson–Crick base-pair planes, which create patterns of hydrogen bond donor and acceptor sites within the plane of the bp. Thus the floor of the grooves of the DNA helix would differ for individual base-pair sequences with respect to their patterns of hydrogen-bond donors, hydrogen-bond acceptors and sites available for hydrophobic interactions (e.g. the C5 methyl group of thymine). These differences form the basis, in part, for the sequence-selective binding of certain ligands (Steitz, 1990). The microheterogeneity of the DNA results from localized, sequence-dependent changes in the shape of the helix, which occurs to limit steric interactions between the purines of adjacent bps (Calladine, 1982).

4.3.2.2 A-form DNA. B-DNA transforms into an A-form helix as the relative humidity of its environment decreases to 75% and the NaCl concentration drops below 10%. A-DNA helices are short and fat with the two DNA strands wrapped around the helix axis, and with the bps and backbone far away from the helix axis. The A-form of DNA is a right-handed helix with 11 base pairs per complete turn and a helical pitch of 2.8 nm (Dickerson *et al.*, 1982). The most pronounced feature of this structure is the 20° tilting of the base-pair planes and their net displacement away from the central axis (Figure 4.5A). The A-form helix adopts a C3′-endo (N-type) sugar pucker conformation as opposed to the C2′-endo conformation present in B-DNA. These structural features result in a deep and narrow major groove and a very shallow and wide minor groove (with reference to B-DNA). In addition to being a helical conformation adopted by double-stranded DNA, the A-form helix is also adopted by RNA-DNA hybrids (Wang *et al.*, 1982) and double-stranded RNAs (Arnott and Selsing, 1974) as a result of the C2—OH substituent which, due to steric interactions, forces the sugar to assume C3′-endo conformations.

4.3.2.3 Z-form DNA. While the most prominent feature of Z-DNA is its ability to adopt a left-handed helical structure, this form of DNA is not simply the mirror image of the B-form or A-form (Rich *et al.*, 1984). The Z-form helix is elongated and slender (Figure 4.5C) with 12 bps per complete turn and a helical pitch of ~4.5 nm. Z-DNA contains a wide and shallow major groove with a narrow and extremely deep minor groove. Z-Form DNA also adopts sugar puckering and *N*-glycosidic torsion angles that alternate between C2′-endo and C3′-endo and between *anti*- and *syn*-respectively, making the actual repeating unit for the Z-form helix two bps in contrast to the single base-pair repeat unit of the A- and B-form DNAs. This alternating pattern, in combination with the 180° rotation of the bases about the glycosidic bond, results in a left-handed helix with a phosphate backbone that appears to zigzag around the helical structure, hence the name Z-DNA. The transition of DNA into Z-form generally requires high salt (3–4 M NaCl) and an alternating pyrimidine-purine sequence. Z-DNA can form in regions of alternating purine-pyrimidine sequence, i.e. $(GC)_n$ sequences form Z-DNA most easily. $(GT)_n$ sequences also form Z-DNA but they require greater stabilization energy for formation than $(GC)_n$. $(AT)_n$ generally does not form Z-DNA since it easily forms cruciform. Additionally, methylation of the C5 position of cytosine and negative torsional stress in supercoiled DNAs also appears to facilitate Z-form helix formation. Z-DNA helices have been demonstrated to exist in polytene chromosomes of *drosophila* by fluorescent antibodies (Nordheim *et al.*, 1981). Structurally, Z-DNA would serve to relieve the strain on DNA when it becomes negatively supercoiled and packed into chromosomes (Klysik *et al.*, 1988). Z-DNA specific binding proteins have been discovered in chromatin, and it has been established that Z-DNA occurs in metabolically active nuclei. Its formation is dependent on DNA torsional strain and increases with transcription (Wittig *et al.*, 1992).

4.3.2.4 Other forms of DNA. Numerous other subtle variations in the shape of the DNA double helix, in specialized situations, may have biological relevance (Saenger, 1984). C-DNA forms in fibers at 57–66% humidity and has 9.3 bp per turn. D-DNA is a structure with a helix repeat of 8.5 bp per turn. Runs of poly(dA)·poly(dT) are believed to adopt a D-like helix. T-DNA has a helix repeat of 8 bp per turn. It is purified from bacteriophages T2, T4 and T6 and is different from most DNA. The cytosine residues in T-DNA contain 5-hydroxymethyl group, which also is glucosylated. This change is reflected in the different shape of the helix of bacteriophage T-DNA. In ordinary double-helical DNA, an axis of symmetry relates to the two strands. This is a pseudo C2 axis since it applies only to the backbones and not to the bases themselves. C2 symmetry implies that a structure is composed of two identical parts. An axis of rotation can be found that interchanges these two parts by a 180° rotation. This axis is called a C2 axis. Pseudo C2 symmetry means that some aspects of a structure can be interchanged by a rotation of 180°, while other aspects of the structure are altered by this rotation.

4.3.3 Alternative structures of DNA

The structure of DNA comprises a series of flat aromatic heterocyclic bases attached to a hydrophilic, negatively charged sugar-phosphate backbone. The backbone is relatively flexible and can relax to accommodate structural alternation caused by variation in base–base interactions. Three main kinds of interactions govern most of the structural changes observed in DNA:

1. The charged phosphate groups of the backbone are mutually repulsive, and this sets limits on the extent of structural variation that is possible. These electrostatic

interactions normally require screening by metal cations, and ion interactions are critical to the formation of many structures.

2. Hydrogen bonding between bases is the driving force for the formation of mutistranded structures.
3. The bps exert short-range attractive interactions on each other, leading to stacking interactions between the flat surfaces.

These properties lead to the adoption of the normal double-stranded structure of DNA comprising the charged backbone on the outside and the stacked bps on the inside. Adjustments in the structural relationship between adjacent bps can be achieved through several operations. These include:

- changes in the local helical twist;
- translation along the long base-pair axis (slide);
- an overall tilting of the mean base-pair plane (roll); and
- the opposite rotation of paired base along their long axis (propeller twist).

Examples of some known alternative structures of DNA duplexes are illustrated in Table 4.6. Importantly, these subtle differences in structure may be another factor that contributes to the selective recognition of the DNA helix by proteins or small molecules. They may also determine the extent to which the structure may be altered as follows:

4.3.3.1 Sequence dependent modification to B-DNA. *The helical repeat of double-stranded DNA*: Stretch of oligo dA·oligo dT adopt a structure called B′ in which there is a pronounced propeller twist of the A·T base pairs. This structure, which is the basis of sequence directed curvature of DNA (Dickman and Wang, 1985; Wu and Crothers, 1984), is relatively overwound. By contrast, alternating adenine-thymine tracts, $(dApT)_n$ appear to be of particularly low torsional stiffness and are capable of being either underwound (McClellan *et al.*, 1986) or overwound (McClellan and Lilley, 1991). Certain

TABLE 4.6 Examples of miscellaneous known DNA duplex structures. Some miscellaneous known DNA duplex structures accessible at NDB (http://ndbserver.rutgers.edu) are illustrated

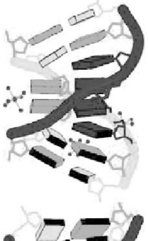
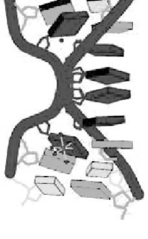
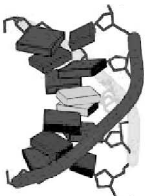
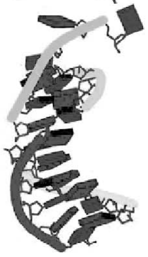
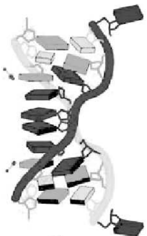
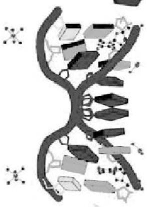
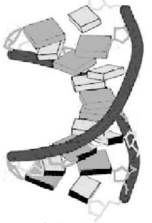
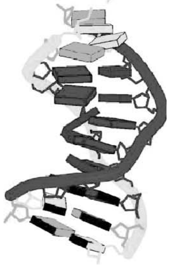
Duplex structure	Model	NDB ID
Bending duplex with a bulge		UD0027
Duplex bulges		UD0034

TABLE 4.6 ***continued***

Duplex structure	Model	NDB ID
Duplex with Hoogsteen base pairing		UD0035
Duplex with flipped-out bases		UD0049
Duplex with overhanging bases		UDIB70
Duplex with base intercalated		UD0032
Duplex with 5-methylC		UD0045
Duplex with thymine dimer		UD0055

(G + C)-rich sequences exhibit a propensity for adopting A-DNA structure. The A-structure is underwound relative to B-DNA with a helical repeat of 11–12 bp/turn.

The handedness of DNA: The most famous example of left-handed DNA is the Z-conformation adopted by alternating pyrimidine-purine sequences, especially by d(CpG)$_n$ repeats (Wang *et al*., 1979). The repeating unit of Z-DNA is a dinucleotide, with alternation of sugar pucker and glysosyl torsion angle, giving a pronounced zigzag appearance. In Z-DNA, the minor groove is extremely deep, while the major groove is convex. The exposure of bases in the major groove leads to chemical reactivities (Herr, 1985) and highly immunogenic (Lafer *et al*., 1981). Segments of Z-DNA can form within a DNA molecule that is predominantly right-handed (Nordheim and Rich, 1983). In this case, interfacial B—Z junctions are created, which are susceptible to attack by a number of enzyme and chemical probes (Galazda *et al*., 1986; Singleton *et al*., 1984), indicative of a perturbed conformation.

Alternation to Watson–Crick base pairing: These are normal DNA structures with complementary bps of adenine with thymine, and guanine with cytosine, which are the most stable bps thermodynamically (Breslauer *et al*., 1986). However, single mispaired bases can be incorporated into the overall geometry of B-DNA with very little disruption in many cases. Rearranged base-pairing may involve hydrogen bond donor and acceptors other than the normal ones, such as the use of guanine N7and O6 and of adenine N7 and N6 (Hoogsteen, 1963). These positions are the basis of altered base-pairings in the accommodation of a third strand in the H-triplex structure (Johnston, 1988) and the formation of guanine tetrads in tetraplexes formed from telomere-related sequences (Laughlan *et al*., 1994; Sundqluist and Klug, 1989).

Deformation of the helix axis: The deformation can be discontinuous as a kink, hairpin and pseudoknot or the curvature may be less localized. An example of the kink is the kinking introduced by base bulges in DNA or RNA (Rice and Crothers, 1989). The most extreme case is the pronounced curvature of DNA associated with oligoadenine tracts that are phased with the helical repeat (Dickman and Wang, 1985; Hagerman, 1985; Wu and Crothers, 1984), but lesser distortions are clearly found in many sequences and truly straight segments of DNA are probably rare. In addition to intrinsic curvature, certain sequences may exhibit enhanced bendability in response to protein binding. The axis of double-stranded DNA can be curved by particular sequences, especially those found in kinetoplast circles (Dickman and Wang, 1985; Hagerman, 1985). This can be observed readily by the resulting anomalously slow electrophoretic mobility (Dickman and Wang, 1985; Hagerman, 1985). The most pronounced sequence-directed curvature results from runs of oligoadenine·oligothymine having lengths of four bps or greater, which are phased with the helical repeat.
e.g. curvature center:

CCC**AAAAA**TGTC**AAAAAA**TAGGC**AAAAAA**TGCC**AAAAA**TCCC
GGG**TTTTT**ACAG**TTTTTT** ATCCG**TTTTTT** ACGG**TTTTT**AGGG

The structure, generally termed B′, has a narrow minor groove and is characterized by a large propeller twist (average 20°) of the A·T base pairs.

Helical junctions: Junctions between helical segments are formed as intermediates in a number of DNA rearrangements (e.g. three-way junction, bulged three-way junction and four-way junction). The most important of these is the four-way junction known as the Holliday junction of genetic recombination (Holliday, 1964; Orr-Weaver *et al.*, 1981). When the counter-ion screening by salt is high, the junction can undergo the folding process. The perfect three-way junction comprises three helices linked through the covalent continuity of the strands, with no unpaired bases.

The number of strands: Standard DNA is double-stranded. However, the width of the major groove in B-DNA is sufficient to accommodate a third strand, and thus triplex DNA can be formed, mainly by oligopurine-oligopyrimidine sequences (Johnston, 1988). Four-stranded DNA has been described for guanine-rich sequences, based on the guanine tetrad (Sundquist and Klug, 1989).

Relative orientation of strands: Normal DNA is a double helix in which the two strands have an antiparallel orientation. In H-triplex structure, the incoming oligopyrimidine strand is oriented parallel to the oligopurine strand of the duplex and antiparallel to the complementary oligopyrimidine strand (Johnston, 1988). Both parallel (Laughlan *et al.*, 1994; Sen and Gilbert, 1990; Wang *et al.*, 1991) and antiparallel (Sundquist and Klug, 1989) form of the guanine tetraplex have been observed.

Defined ordered sequences: Within the non-coding regions and on occasion within coding regions, there are defined ordered DNA sequences (dosDNA) that contain various symmetry elements, inverted repeats, mirror repeats and direct repeats. In the case of an inverted repeat, not only are the two strands of the DNA complementary to each other, but each single DNA strand is self-complementary within the inverted repeated region. In this configuration the sequence of DNA in either strand can form a local hydrogen bonded cruciform (Beerman and Lebowitz, 1973; Lilley, 1989). A mirror repeat has AN identical bp in one strand equidistant from a center of symmetry (Figure 4.7). However, these pairs cannot form Watson–Crick hydrogen bonds because the complementary bps are parallel, not antiparallel, in their orientation. Certain sequences with mirror repeat symmetry can form triplex structures (Mirkin *et al.*, 1987). Direct repeats are regions of DNA in which a particular sequence is repeated or duplicated. The repeat can be either adjacent to or located at some distance from the first repeat. A number of non-B-DNA conformations can form within dosDNA sequences, including cruciform, intramolecular triplex DNA and slipped mispaired structures.

4.3.3.2 Locally disrupted base pairing can take a number of different forms.

Single-base mismatches refer to single non-Watson–Crick base opposition flanked by normal base pairing. The disruption to the overall geometry of the double helix is quite

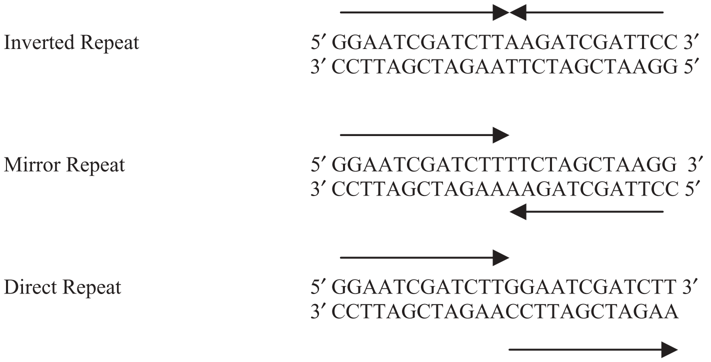

Inverted Repeat
5′ GGAATCGATCTTAAGATCGATTCC 3′
3′ CCTTAGCTAGAATTCTAGCTAAGG 5′
Mirror Repeat
5′ GGAATCGATCTTTTCTAGCTAAGG 3′
3′ CCTTAGCTAGAAAAGATCGATTCC 5′
Direct Repeat
5′ GGAATCGATCTTGGAATCGATCTT 3′
3′ CCTTAGCTAGAACCTTAGCTAGAA

Figure 4.7 Symmetry elements in DNA. The organization of three different types of symmetry elements in DNA is shown. They are inverted repeats, mirror repeats and direct repeats. The arrows above and below the base sequences show the organization of symmetrical complementary sequences in DNA

small with the mismatched bases remaining stacked into the helix and hydrogen bonded. However, the local structure of the mismatched leads to an enhanced chemical reactivity (Cotton *et al.*, 1988).

Multiple-base mismatches refer to a group of consecutive non-Watson–Crick base oppositions. This is similar to a locally melted bubble or internal loop. Multiple mismatches may introduce a point of flexibility WITH the bases exhibit strong reactivity toward enzyme and chemical probes (Bhattacharyya and Lilley, 1989).

Bulged bases may occur with one or more extra bases that are unopposed on the complementary strand of duplex. Base bulges introduce a pronounced kinking of the helix axis, which can readily be observed as an electrophoretic retardation of the bulged molecule (Bhattacharyya and Lilley, 1989; Rice and Crothers, 1989).

Examples of some alternative (unusual) DNA structures are described below.

Slipped, mispaired DNA. Slipped, mispaired DNA structures can form in regions with direct repeat symmetry. To form a slipped, mispaired structure, the entire region must unwind and one strand of one copy of the direct repeat must pair with the complementary strand of the other copy of the direct repeat. The slipped, mispaired DNA (smp-DNA) has two isomers; one isomer has loops composed of the 5′ direct repeat in both strands, whereas the other has loops composed of the 3′ direct repeats in both strands. Because this structure results in the unwinding of the DNA duplex, DNA supercoil will be lost on formation of smp-DNA. Regions of eukaryotic DNA with direct repeat symmetry that are sensitive to S1 nuclease (McKeon *et al.*, 1984) have suggested the existence of smp-DNA. Biologically, smp-DNA is important in spontaneous frameshift mutagenesis (a frameshift mutation is one that changes the reading frame of DNA). It is known that spontaneous deletion or addition mutations can occur within runs of a single base (Streisinger *et al.*, 1966). Since the genetic code is read as triplets, adding or deleting a single base shifts the reading frame of all bases downstream from the point of mutation. This event will result in an mRNA that encodes amino acids that are different from those present in the wild-type protein downstream from the frameshift mutation. Large deletion and duplication mutations occur between direct repeats that can form slipped structures during DNA replication. These direct repeats can be adjacent to one another or separated by as many as hundreds of bps. The nature of the intervening DNA between the direct repeats is also important. If DNA between the direct repeats contains palindromic symmetry, the formation of a hairpin arm can stabilize the misalignment and increase the frequency of deletion (Williams and Müller, 1987).

Cruciform structures in DNA. Cruciform structures can be regarded as a special example of a four-way junction (Murchie *et al.*, 1992). They are formed from inverted repeat sequences, when the two strands each form intrastranded hairpin structures. The inverted repeat orientation has also been called a ‘palindrome’. Since the DNA has polarity and is read by RNA polymerase in the 5′ → 3′ direction, the term palindrome can be used for perfect inverted repeats. Inverted repeat regions that are not completely symmetrical or that have a center region without inverted repeat are not considered palindromes (sometime termed imperfect palindromes or quasi-palindromes). Often inverted repeats occur near putative control regions of genes or at origins of DNA replication. To form a cruciform, the interstrand hydrogen bonds in the inverted repeat must melt and intrastrand hydrogen bonds must form between complementary bases in each single strand. On cruciform formation, there must be a loop of 3–4 unpaired bases at the tip of the hairpin, resulting in a loss of hydrogen bonding and base stacking interactions that provide stability to the linear double helix. Cruciforms are stable and thermodynamically favored in negatively supercoiled DNA because, when cruciform are formed, negative supercoils are relaxed. The formation of cruciforms results in the relaxation of one negative supercoil for every 10.5 bp of the inverted repeat that participates in cruciform formation. The stability of cruciform results from a loss of the free energy inherent in supercoiled DNA. Thus the stability of cruciforms is proportional to their length. Cruciforms are stable, even in relaxed DNA, once they are formed, because removal of a cruciform in relaxed DNA requires the introduction of one negative supercoil for each 10.5 bp of the cruciform arm converting back into a linear duplex.

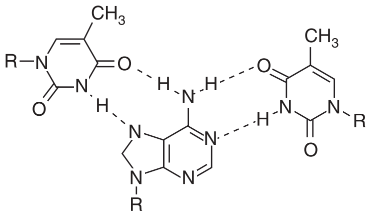
Two kinetic processes are involved in the extrusion from unperturbed DNA to the cruciform structure (Lilley, 1985). The formation of cruciform DNA in buffer that approximates physiological ionic strength, typically 50 mM NaCl is termed S-type mechanism (salt dependent). The S-type transition is dependent on supercoiling, temperature, ionic conditions and divalent cation (Sullivan and Lilley, 1987). Under physiological conditions (S-type), the rate of cruciform formation is dependent on the base composition at the center of symmetry. There is a local opening at the center of the inverted repeat, formation of aprotocruciform, and finally branch migration to the fully extruded cruciform. S-type kinetics is affected by the sequence and methylation state at the center of the inverted repeat (Murchie and Lilley, 1987). A second pathway by which cruciforms can form in solution lacking salt is term C-type mechanism (Lilley, 1985; Lilley, 1989). If the sequence context of the inverted repeat is very (A + T)-rich and the ionic strength is low, there appear to be a much larger cooperative opening of the entire region along with a single step formation of the cruciform, the C-type mechanism (Bowater *et al.*, 1991). The two types differ in three ways:

1. The C-type cruciform transition occurs at low temperature <30°C in contrast to S-type formation, which generally requires higher temperature.
2. C-Type cruciform formation requires very low ionic strengths. Counter ions stabilize the DNA double helix, and in a solution of very low ionic strength, the DNA double helix is much more readily unwound.
3. C-Type cruciform formation has a large energy of activation of about 180 kcal/mol, while S-type formation is about 40 kcal/mol.

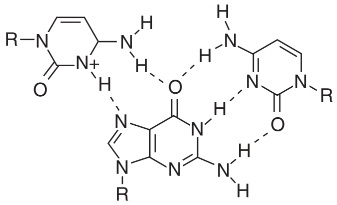
The prevalence of inverted repeats in bacterial, eukaryotic and viral DNA implies a biological role for cruciforms. However, in the linear form, inverted repeats play important biological roles as binding sites for dimeric proteins. It is believed that cruciforms could play biological roles in transcription or DNA replication. Many short inverted repeats, ranging in size from 4 to 20 bp, represent the binding sites for specific proteins. Inverted repeats are frequently associated with origins of DNA replication, such as hairpins in single-stranded bacteriophage genome, eukaryotic viruses and several plasmids.

Triplex DNA. A normal Watson–Crick paired helix of poly(dA)·poly(dT) can form hydrogen bonds with an additional strand of poly(dT), yielding a $(dA)(dT)_2$ triple helix (Felsenfeld and Miles, 1967; Arnott and Selsing, 1974), written as (*dT*)·(dA)·(dT) with the third strand in *italics*. Triplex formation can result by uptake of a third strand in the major groove of a duplex (Arnott and Selsing, 1974). This can occur by the uptake of an exogenous third strand or by rearrangement of a duplex structure to generate an intramolecular triplex. Triple helices also formed in RNA strands (e.g. $(rA)(rU)_2$). Triple-stranded DNA is formed by laying a third strand into the major groove of DNA which, in the triplex form, may adopt an unwound B-DNA-like conformation. Complementary hydrogen bonding interactions are responsible for the specificity of the third strand interaction. Since the

Watson–Crick base-pairing surfaces are already involved in hydrogen bonding within the duplex, the third strand must hydrogen bond to another surface of the duplex. The third strand participates in a Hoogsteen base-pairing scheme in triple-stranded DNA(*T*AT, *C*GC, *G*GC, and *A*AT).



T·A·T triplex



C·G·C$^+$ triplex

The central strand of the triplex must be purine rich since a pyrimidine does not have two hydrogen bonding surfaces with more than one hydrogen bond. Thus triple-stranded DNA requires a homopurine·homopyrimidine region of DNA. If the third strand is purine rich, it forms reverse Hoogsteen hydrogen bonds in an antiparallel orientation with the purine strand of the Watson–Crick helix. If the third strand is pyrimidine rich, it forms Hoogsteen bonds in a parallel orientation with the Watson–Crick paired purine strand.

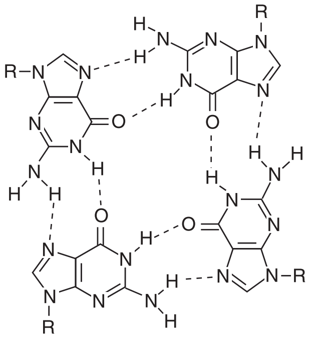
The formation of an intramolecular triplex could form within a single homopurine·homopyrimidine duplex DNA region in supercoiled DNA (Lyamichev *et al.*, 1986; Mirkin *et al.*, 1987). This is of interest because many sequences in the human genome have the potential to form intramolecular triplex structures. These sequences are commonly associated with regulatory regions of genes. In addition to requiring a continuous strand of purine bases, the homopurine·homopyrimidine region must contain mirror repeat symmetry (Mirkin *et al.*, 1987). The mirror repeat is a region of DNA that has the same base sequence reading in both the 3′ and the 5′ directions (from a center point) in one strand of DNA. There are different ways in which a homopurine·homopyrimidine mirror repeat sequence can fold into an intramolecular triplex (H-DNA). The most common structure is the *Py*·Pu·Py configuration in which half of the pyrimidine strand pairs as the third strand and the complementary strand of this region remain unpaired. The two different *Py*·Pu·Py structural isomers that can form, the Hy5 or Hy3 isomers, correspond to structures in which the 5′ or 3′ half of the pyrimidine strand pairs as the third strand (Htun and Dahlberg, 1989). Similar structures Hu5 and Hu3 can form, in which the third strand is the purine strand.

There are two general requirements for intramolecular triple stranded DNA formation. First, a homopurine·homopyrimidine region is preferred to provide a continuous purine central strand. Second, the region should contain mirror repeat symmetry. The third strand-pairing rule (Letai e*t al.*, 1988) defines that the central strand must be a purine. A central guanosine can form a Hoogsteen bp with C, G or I (inosine). Adenosine can pair with T, A or I. The majority of intramolecular triplexes studied involve the regions of polypurine·polypyrimidine sequence that have mirror repeat symmetry. Typically the polarity of the third strand is reversed for a *Pu*·Pu·Py triplex and a *Py*·Pu·Py triplex helix, which contain reverse Hoogsteen and Hoogsteen bps respectively.

The widespread occurrence of polypurine·polypyrimidine tracts in eukaryotic DNA suggests that these sequences may have a biological function. Analysis of eukaryotic sequence databases reveals thousands of polypurine·polypyrimidine tracts, many with the potential for triplex formation. These polypurine regions of DNA can potentially influence biology in several ways. They could provide binding sites for regulatory proteins, influence nucleosome positioning or form intramolecular triplex structures. The ends of telomers of linear eukaryotic chromosomes may exist as triplex structures. A triplex at the end of chromosomes may contribute to the stability of ends, preventing digestion by exonucleases or participation in genetic recombination.

Intramolecular triplexes, or H-DNA, have been observed in supercoiled DNA at low pH in which half of the pyrimidine strand of the oligopurine·oligopyrimidine forms the third strand of a triplex structure (Lyamichev *et al.*, 1986). Thus the pyrimidine strand forms a kind of hairpin structure, while half of the purine strand is left formally single-stranded. Sequence variation showed the requirement for an oligopurine sequence with a mirror repeat (Hanvey *et al.*, 1988; Mirkin *et al.*, 1987), as expected for the H-structure.

Tetraplex DNA. Sequences containing repeated runs of oligoguanine can adopt conformations based on association between guanine bases (Sundquist and Klug, 1989). Such sequences are found in chromosome telomers, which are generally based on a repeated unit of either $T_{2\text{-}4}G_4$ or $T_{2\text{-}4}AG_3$, with a single stranded purine-rich 3′-overhang of 12–20 bases (Zakian, 1989). These structures are based on the inter- or intramolecular assembly of four strands, held together by the formation of tetrads of guanine bases that are hydrogen-bonded between N1 and O6 and between N2 and N7 in cyclic manner.



Guanine tetrad

Tetraplex formation may occur either as a parallel intermolecular association of four strands (Sen and Gilbert, 1990), as an antiparallel association of either two hairpin-forming strands (Sundquist and Klug, 1989) or a single strand that is folded back to comprise three loops. All glycosyl torsion angles are *anti* in the parallel structure (Aboul-ela *et al.*, 1994), while in the antiparallel structure, *syn* and *anti* glycosyl angles alternate along the chain (Wang *et al.*, 1991). The parallel tetraplex structure has also been observed in RNA (Kim *et al.*, 1991), and RNA tetraplex formation has been suggested to be important in retroviral genome dimerization (Sundquist and Heaphy, 1993).

4.4 SUPERCOILING AND TERTIARY STRUCTURE OF DNA

4.4.1 DNA topoisomers

In most organisms, DNA exists in a supercoiled form. Supercoiled DNA can exist in a wide variety of topological conformations with variations in twisting and writhing. A simple plasmid molecule purified from bacterial cells will exist as a naturally occurring covalently closed circular DNA molecule (sometimes called cccDNA) that is negatively supercoiled (form I DNA). Supercoiled DNA appears as a compacted molecule in which the DNA helix has wound about itself. If a covalently closed DNA molecule contains a single nick in one of the strands, the nick provides a swivel and supercoils are lost. Nicked DNA (form II DNA) does not contain the supercoils or supertwists (writhes) of the helix and will appear as a circular ring, which is indistinguishable under microscope from the cccDNA that contains no supercoils (relaxed DNA). DNA that contains breaks in both phosphate backbones at the same point (or nearly the same point) along the helix axis will for a linear DNA molecule.

One topological property of a DNA molecule is its linking number, L. The linking number (integer) is defined as the number of times one strand crosses the other when the DNA is made to lie flat on a plane. The linking number cannot change unless the phosphodiester backbone is broken by chemical or enzymatic cleavage. Although L cannot change, the number of twists or turns of the double helix, as well as the number of supercoils or writhes of the double helix, can change. The topology of DNA is described by the simple equation:

$$L = T + W$$

where L, the linking number is as defined; T is the number of helical turns in the DNA; and W is the writhing number of DNA, which describes the supertwisting or coiling of the helix in space. Although the linking number is invariant and must be an integer, the number of twists (turns) can vary in positive and negative increments with offsetting negative and positive changes in the writhe number. A simple explanation of DNA supercoiling has been provided (Scovell, 1986). For example, the introduction of a negative supertwist (a right-handed coil) into the DNA molecule with $L = 18$ changes the value of W by −1 ($W = -1$). Since L cannot change ($L = 18$), T must increase by +1 ($T = 19$). Conversely, a decrease in the turn (twist) number by 1 ($T = 17$) would require a compensating introduction of a left-handed or positive supertwist ($W = +1$). DNA is a dynamic molecule that can exist in myriad states with different values of helical turns (twists) and supercoils (writhes). The change in twists and writhes of a supercoiling sample can be estimated by two-dimensional gel electrophoresis (Howell *et al.*, 1966). The structural heterogeneity resulting from the flexibility in the winding and the bendability of the DNA helix allows the supertwisting or coiling of the helix. These supertwisted/coiled DNA topoisomers include:

Relaxed DNA. The DNA helix in solution will adopt a preferred helical repeat that is function of the base composition. In linear DNA, in which the ends of the molecule are free to rotate, the DNA will adopt this preferred helical repeat of 10.5 bp per turn in solution. The helical repeat also exists in nicked DNA in which the nick provides a swivel whereby one strand can rotate about the other. The preferred helical repeat of a linear DNA molecule represents the lowest energy form of the molecule. When this state of helical twist exists in a covalently closed molecule, the molecule is relaxed and contains no supercoils. In the relaxed DNA, the linking number equal the twist number ($L = T$) and $W = 0$. The linking number of relaxed DNA, L_0 is defined as:

$$L_0 = N/10.5$$

where N is the number of bps in the DNA molecule and 10.5 refers to the helical repeat. The L_0 value is the same for the DNA molecule in a linear, nicked or covalently closed relaxed form. Thus the DNA molecule is relaxed if $L = L_0$.

Negatively supercoiled DNA. DNA is said to be negatively supercoiled when $L < L_0$. Negatively supercoiled DNA is underwound because it has fewer helical turns than the molecule would contain as linear or relaxed molecule. This underwinding in the number of helical turns results in more bps per helical turn. This results in a decrease in the angle of twist (or the rotation per residue) between adjacent bps. Therefore underwinding creates torsional tension in the winding of the DNA double helix. In the molecule (e.g. 210 bp) with $L = 18$, there are only 18 helical turns. This will change the average rotation per residue from 34.29° [(20 × 360°)/210] to 30.86° [(18 × 360°)/210]. Energetically, this represents an unfavorable winding of the DNA double helix. The driving force for forming the supercoils may be viewed as the underwinding in the number of helical turns and the resulting torsional strain inherent in the decreased angle between adjacent bases in the winding of the double helix.

Most DNA in its natural state, including chromosomal and plasmid DNA in bacteria as well as DNA in human cells, is negatively supercoiled. Plasmid and circular bacterial chromosomes are topologically closed and thus have a defined linking number. This property of being a topologically closed system is essential for DNA supercoiling. The *E. coli* chromosome exists as a large (2.9×10^6 bp) closed circle that constitutes one large topological domain. A topological domain is defined as region of DNA bounded by constraints on the rotation of the DNA double helix. The large circle is subdivided into about 45 independent topological domains *in vivo*. Eukaryotic chromosomes are believed to exist as long linear molecules. To exist in a supercoiled state, linear DNA must become organized into one or several topological domains. Independent loops are believed to be formed by the interaction of specific regions of DNA with defined proteins that attach to the nuclear matrix. DNA could become supercoiled in living cells in number of ways:

- through the action of DNA gyrase/topoisomerase;
- by wrapping DNA into a negative supercoil around a protein; and
- by the act of transcription.

Positively supercoiled DNA. Most DNA isolated from natural sources is negatively supercoiled. However, DNA can exist in a positively supercoiled form when it has a greater linking number than relaxed DNA, i.e. $L > L_0$. Positively supercoiled DNA is overwound in terms of the number of helical turns. This overwinding creates a situation in which there are fewer bps per helical turn, resulting in an increase in the winding angle between adjacent bps. This creates torsional tension in the winding of the DNA double helix. Overwinding the 210-bp DNA molecule by two turns to $L = 22$ creates a situation in which the average rotation per residue changes from 34.29° in relaxed DNA to 37.71° [(22 × 360°)/210]. The tension in the winding of the helix is relieved by the positive supercoiling of the DNA forming a left-handed (counterclockwise) supercoil. Positively supercoiled DNA has been isolated from a bacteriophage-like plasmid from a *Sulfolobus* species (an archebacterium living at high temperature and low pH) (Nadal *et al.*, 1986). Positively supercoiled DNA would resist unwinding of the helix by heat and acid.

The role of supercoiling in gene expression. DNA supercoiling will affect the rotational relationship between the two important region of the promoter, and the ability of the region to breathe. It can also differentially affect individual kinetic steps in transcription initiation. Different promoters have been shown to be tuned to work best at different levels of DNA supercoiling. The organization of the chromosome into topological domains allows for the possibility that different regions of the chromosomes are differentially supercoiled. It has been realized that the state of supercoiling of the DNA is an important factor in triggering the initiation of the replication process. DNA, through its shape and torsional flexibility, works in concert with the replication proteins to begin the process of replication.

Toroidal coils, knots and catenanes. Supercoils in DNA need not physically exist as interwound supercoils. The negative supercoils can exist as left-handed toroidal coils, which topologically satisfy the requirement for W. Although in a toroidal coil the helix does not cross itself in the fashion of an interwound supercoil, it does cross itself in the plane of the toroidal coil. The organization of DNA in nucleosomes in eukaryotes involves the toroidal coiling of the DNA around proteins.

A circular DNA molecule can fold into a knot. This requires the introduction of a double-stranded break and the wrapping of the DNA around itself before resealing the double-stranded break. Two circular DNA molecules can also link together to form catenanes (or interlocked rings). Knots and catenanes can have a negative or positive sign, depending on the order of the strand crossing or nodes (places where one helix crosses the other). By convention, if an arrow drawn along the top strand is rotated <180° in a clockwise direction to align with the arrow drawn in the same direction along the bottom double helix of the node, then the node is negative. If the top arrow must be rotated counterclockwise <180° to align with the bottom arrow, the node is positive. DNA knots and catenanes are found *in vivo* as the products of certain topoisomerases and enzymes involved in the site-specific recombination (Wasserman and Cozzaralli, 1986).

4.4.2 Superhelical density and energetics of supercoiling

Superhelical density, σ, is a term frequently used and is defined as the average number of superhelical turns per helical turn of DNA:

$$\sigma = 10.5\tau/N$$

where τ is the number of measurable supercoils (e.g. by electron microscopy, two-dimensional agarose gels and solution titration methods), N is the number of bps in the molecule, and 10.5 represents the average number of bps per turn. The specific linking number difference σ_{sp} is a term that is also used to describe the level of superhelical density:

$$\sigma_{sp} = (L - L_0)/L_0$$

σ_{sp} refers to the inherent specific linking difference and the value of σ_{sp} can be different from that of σ.

The free energy of supercoiling is proportional to the square of the linking number difference in the DNA:

$$\Delta G = (1100RT/N)(L - L_0)^2$$

where R is the gas constant and T is the temperature in degrees Kelvin. N is the number of bps in the DNA molecule. Supercoiled DNA contains a large amount of free energy that can be used to drive biological reactions. Any reactions that lower the free energy of supercoiled DNA will be thermodynamically favored. The biological processes of DNA replication and transcription require an input of energy to open or unwind the DNA double helix to expose the chemical identity of the bases at the center of the helix. Some of the energy for these processes comes from the free energy of supercoiling. The unwinding of the DNA double helix is thermodynamically favored in supercoiled DNA.

4.5 CLASSIFICATION AND STRUCTURES OF RNA

4.5.1 Structures of RNA

RNA chains are made up of nucleotides, which have 3′ → 5′ phosphodiester linkages. The 3′ → 5′ linkage in RNA is thermodynamically less stable than the unnatural 2′ → 5′ linkage, which might therefore have had an evolutionary role. A rare example of such a polymer is produced in vertebrate cells in response to viral infection. Such cells make a glycoprotein called interferon, which stimulates the production of an oligonucleotide synthetase. This polymerizes ATP to give oligoadenylates with 2′ → 5′ phosphodiester linkages such as $(2' \rightarrow 5')(A)_n$ ($n = 3 - 8$), then activates an interferon induced ribonuclease (RNase N) whose function seems to be to break down the viral mRNA. The 2′ → 5′ ester linkage is also a key feature of self-splicing RNA. Qualitatively, the conformation of single-stranded RNA folds so that its complementary sequences form double helixes to the maximum extent possible, with the resulting hairpins or stem-loops constituting its secondary structure. The secondary structure of RNA is often independently stable. Interactions between stem-loops largely determine the overall fold or tertiary structure. The tertiary structure of RNA is thus composed of secondary structural motifs that are brought together to form modules, domains and the complete/global structure.

RNA differs from DNA by having an additional hydroxyl at the 2′-position of the sugar. This has two major implications that distinguish the chemical and physical properties of RNA and DNA. The 2′-OH makes RNA unstable with respect to alkaline hydrolysis. Thus RNA is a molecule intrinsically designed for turnover at the slightly alkaline pH normally found in cells, while DNA is chemically far more stable. The 2′-OH also restricts the range of energetically favorable conformations of the sugar ring and the phosphodiester backbone. This limits the range of conformations of the RNA chain, compared to DNA, and it ultimately restricts RNA to a much narrower choice of helical structures. However, the 2′-OH can participate in interaction with phosphates or bases that stabilize folded chain structures. As a result RNA can usually attain stable tertiary structures (ordered, three-dimensional, relatively compact structures) with more ease than the same corresponding DNA sequence.

The presence of the 2′-hydroxy group in RNA hinders the formation of a B-type helix but can accommodate with an A-type helix. Double-stranded RNA forms A-type helices invariably. At low ionic strength, A-RNA has 11 bp per turn in a right-handed, antiparallel double-helix. The sugar adopts a C3′-endo pucker. If the salt concentration is raised above 20%, an A′-RNA form is observed, which has 12 bp per turn of the duplex. Both structures have typical Watson–Crick base pairs that are displaced 4.4 Å from the helix axis and so for a very deep major groove and a rather shallow minor groove. RNA structures are strikingly diverse, reflecting the many biological functions of the polymer. Its functions depend on the ability of this nucleic acid to adopt many different forms, resulting in RNA molecule tailor-made to suit a particular activity, since the precise function of an RNA molecule relies on its ability to maintain a proper 3D structure. Thus the folded domains of RNA molecules (Murphy and Cech, 1993) can be likened to those of globular proteins and do not easily fit into categories like the DNA conformations. Overall, RNA structures are quite distinct. While their global tertiary structures appear diverse, several fundamental units of secondary structure recur, including A-form double-stranded stems, stem-loops and pseudoknots.

RNAs are generally classified according to their cellular functions. Cells contain three major classes of RNA, which participate in protein biosynthesis. They are transfer RNA or soluble RNA (tRNA or sRNA), ribosomal RNA (rRNA) and messenger RNA

TABLE 4.7 Major classes of RNA in *E. coli*

	Relative amount (%)	Number of nucleotides	Sedimentation constant (S)	Molecular mass (kDa)
mRNA	5	75–3000	6–25	$25–1.0 \times 10^3$
tRNA	15	75	4	25
rRNA	80	120	5	36
		1.7×10^3	16	550
		3.7×10^3	23	1.2×10^3

(mRNA) (Table 4.7). mRNA is the template specifying sequences of amino acids for protein synthesis. tRNAs are carriers of amino acids in the active form to the ribosome for peptide bond formations. rRNA in association with ribosomal proteins, provides the site for proteins synthesis and mediates peptidyl transfer in the formation of peptide bonds. Structural classification of RNA can be found at SCOR (http://scor.lbl.gov/). The following sections will illustrate RNA secondary and tertiary structures, through a description of the three main functional classes of RNAs found within a cell.

4.5.2 Transfer RNA

Transfer RNA (tRNA) or soluble RNA (sRNA) molecules have been examined at high resolution (Rich, 1977). Transfer RNAs are relatively compact single strands of nucleic acid of 23–30 kDa in size that contain 74–94 nucleotides (nt). Individual tRNA transport a covalently attached amino acid to the ribosome and facilitate its proper incorporation into a protein sequence, the latter of which is specified by the sequence of nucleotides read as a triplet code in the mRNA. Thus a tRNA molecule contains two key functional domains:

1. the site of covalent attachment for a particular amino acid; and
2. a triplet anticodon base sequence that is complementary to a codon in the mRNA.

Each tRNA is specific for only one particular amino acid. However, the triplet code displays degeneracy and consequently there are often several different tRNA molecules (isoacceptor tRNAs (Sprinzl *et al.*, 1991)) for one particular amino acid, each with a different three-base anticodon. Surprisingly, even though all tRNA molecules are believed to adopt similar tertiary structures, the enzymatic attachment of the correct amino acid on to its cognate tRNA by an aminoacyl-tRNA synthetase is carried out with profound fidelity (Schimmel, 1987). Of particular interest is the determination of those features of a tRNA molecule that lead to the attachment of the proper amino acid.

Transfer RNA from all organisms contains modified nucleosides. Most of them play an important role in the fine tuning of tRNA activity. The presence of a modified nucleoside improves the efficiency of the tRNA in the decoding event, and may affect the fidelity of protein synthesis and codon choice such as the modified nucleoside next to the 3′-side of the anticodon (position 37) and at the Wobble position (position 34). Modified nucleosides in the anticodon region, other than positions 34 and 37, may influence the translational efficiency and fidelity, whereas those outside the anticodon region may stabilize tRNA conformations. Useful information concerning modified nucleosides in RNA is available from RNA modification database at http://medlib.med.utah.edu/RNAmods.

The primary and secondary structure of tRNA molecules are often depicted as a cloverleaf pattern containing double-stranded stems and stems connected to single stranded loops, two recurring elements of RNA secondary structure. Each A-form stem consists of hydrogen-bonded base pairs (G—C and A—U) that occur through self-complementary regions of the tRNA primary sequence. Many tRNAs that exist in nature can contain several constant structural regions that can be described using the cloverleaf motif. At the 5′-termini, each tRNA is phosphorylated and virtually all have a seven base-pair structure referred to as the acceptor stem; it is so named because it provides the point of connection for the appropriate amino acid. The 3′-end of the tRNA, which is the position of amino acid attachment, always terminates in the sequence —CCA_3' and provides a free 3′-OH group for attachment of the amino acid as an activated ester by the cognate aminoacyl-tRNA synthetase. The acceptor stem is also unique in that it often contains a non-Watson–Crick bp such as G—U. In addition to the acceptor stem, the TψC loop contains a five base-pair stem with a highly conserved seven-base loop region that incorporates the modified nucleobase pseudouracil (ψ). All tRNAs also contain:

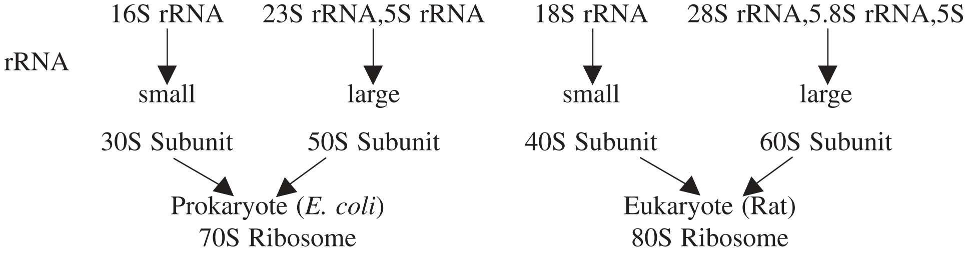
- D-loop that frequently contains the modified nucleobase dihydrouracild (D); and
- variable loop, which differs in nucleotide length between individual tRNAs.

The remaining key structural feature of a tRNA is the anticodon stem, a stem-loop structure that contains the anticodon triplet responsible for hydrogen bonding to the complementary sequence of mRNA and consequent delivery of the proper amino acid to the ribosomal machinery.

The tertiary structure of tRNA resulting from these secondary interactions is distinctly globular in appearance, resembling an L-shape (Figure 4.4). In this structure, one arm of the tRNA, consisting of the acceptor stem and the TψC loop, is folded into an A-form double helix, while the D-loop and the anticodon loop similarly form the other arm of the L. Each arm of this structure is approximately 6.0 nm in length, with the anticodon and the acceptor stem at opposite ends of the molecule. The folded structure of the tRNA is maintained by non-Watson–Crick hydrogen bonding interactions, which serve to cross-link distant regions of the tRNA, locking it into the desired tertiary structure. In addition to hydrogen bonding, stacking interactions within the interior of the molecule are also extremely important in maintaining this structure; over 90% of the bases present are involved in this form of interaction. While this extensive stacking renders most of the tRNA interior inaccessible to solvent, those regions necessary for intermolecular interactions are placed at readily accessible locations.

4.5.3 Ribosomal RNA

Ribosomal RNA (rRNA) is the major component of the ribosome, encompassing approximately two-thirds of the gross weight and forming the functional portion involved in protein synthesis. The other component of the ribosome consists of several proteins that are believed to assist in the maintenance of the required 3D structure of the active rRNA and which presumably have other functions as well.



Examinations of the primary sequence of rRNAs by methods that predict higher-order structures (Zucker, 1989) suggest that they incorporate many of the elements of secondary structure employed by tRNA molecules. Ribosomal RNA molecules contain an array of double-stranded stems and single-stranded loops that resemble the structural elements in the cloverleaf structure of tRNA. In addition, there are interior loops, bulge loops and multibranched loop structures. While prediction methods can generate many plausible forms, analysis of the 16S rRNA from several species consistently gave rise to the four-domain structure. Along with the secondary structure elements discussed, there is increasing evidence to suggest that rRNA (and other form of RNA) often contain a structural unit termed the pseudoknot (Schimmel, 1989). This unique structural element is formed from stem-loop structures that hydrogen bond through the loop to an additional strand of RNA, thereby creating a new stem structure. It is speculated that pseudoknots create unique tertiary structural elements for recognition by other RNAs or proteins and may also provide a conformational switching mechanism between two structural forms of RNA (Dam *et al.*, 1992). The sequence complementarity exists between the regions of the 16S and 23S *E. coli* rRNA. These complementary regions may participate in the quaternary base pairing when a 70S ribosome is formed from 30S and 50S subunits.

The functional roles of rRNAs include the participation in mRNA selection, tRNA binding, ribosomal subunit association, frame-shift suppression, translational proofreading, binding of various factors and peptidyl transferase activity. Messenger RNA is selected by interaction with the 3′-terminus of 16S rRNA in the platform region of the small subunit. tRNA interacts with 16S rRNA in the cleft of the small subunit and around the universally conserved central loop of 23S rRNA (via CCA terminal of tRNA) of the large subunit. This conserved region of rRNA may be involved in the peptidyl transferase activity. Subunit interaction appears to involve elements of both 16S and 23S rRNA.

4.5.4 Messenger RNA

Messenger RNA (mRNA) is transcribed from DNA containing the information to encode protein synthesis and characterized by the following properties:

- Its base composition reflects the base composition of the DNA that specifies it.
- It is heterogeneous in molecular mass/size.
- It should be transiently associated with ribosomes, the sites of protein synthesis.
- It should be dynamics, i.e. high rate of turnover.
- It hybridizes with DNA, which encodes its synthesis.

In stark contrast to tRNA and rRNA, mRNAs are thought to be single-stranded nucleic acids in their biologically active form, although local secondary structures specific for individual mRNA do form. Given its role as a carrier of the genetic code from DNA to the ribosome, the primary activity of mRNA is to permit this code to be read through the formation of complementary hydrogen bonds with the triplet anticodon of a tRNA. To facilitate access to this information, mRNA is likely to remain relatively less structured than other RNAs in the biological milieu. Although the single-stranded form of mRNA predominates, hairpin or stem-loop structure can occur and provide a means of controlling the rate of expression of a particular gene by providing stopping and pausing points for the ribosomal machinery. In addition to the stem-loop features found in the structures, increasing evidence suggests that some mRNAs also incorporate pseudoknots within their ribosomal binding sites (Schimmel, 1989).

4.5.5 Other classes of RNA

Many of regulatory RNAs referred to as small nonmessenger RNAs (snmRNA) or noncoding RNAs (ncRNA) act to regulate gene expression pathways in one of two basic mechanisms:

1. pair–pairing interactions with other nucleic acids such as antisense RNA and miRNA; and
2. acting as aptamers to bind and modify the activity of a protein or protein complex such *E. coli* 6S (184 nt) RNA, human SRA (steroid receptor RNA activator) RNA that modulate transcription (Storz *et al.*, 2005).

Databases of ncRNA can be found at http://www.sanger.ac.uk/Software/Rfam and http://jsm-research.imb.uq.au/rnadb.

Antisense RNA is defined as a short RNA transcript that lacks coding capacity, but has a high degree of complementarity to another RNA (target RNA), which enables the two to hybridize. The consequence is that such antisense, or complementary RNA can act as a repressor of the normal function or expression of the targeted RNA (Eguchi *et al.*, 1991). Such species have been detected in prokaryotic cells with suggested functions concerning RNA-primed replication of plasmid DNA, transcription of bacterial genes, stability of mRNA and messenger translation in bacteria and bacteriophages.

Heterogeneous nuclear RNA (hnRNA) is synthesized in eukaryotic nucleus with varied length. They are large precursors of eukaryotic mRNA (also known as pre-mRNA), which contain coding regions (exons) and noncoding regions (introns). **Small nuclear RNA**s (snRNAs) are a class of RNA molecules found mainly in eukaryotic nucleus. They contain about 60–300 nucleotides and are found in stable complexes with specific proteins forming small nuclear ribonucleoprotein particles (snRNP), which are about 10 S in size. SnRNP are important in the processing of hnRNA into mature mRNA for export from the nucleus to the cytoplasm. The snRNA associated with the nucleolus referred to as **small nucleolar RNA**s (snoRNAs) are likely participated in eukaryotic ribosome biogenesis. SnoRNAs are complexed with specific proteins as ribonucleoprotein particles (snoRNP), which presumably combine with pre-rRNA, ribosomal proteins and nonribosomal proteins involved in rRNA maturation and ribosome subunit assembly (Maxwell and Fournier, 1995).

Small interfering RNA (siRNA) is the small double-stranded RNA (dsRNA) functioning to silence the expression of specific genes at the posttranscriptional level by a pathway known as RNA interference (Harborth *et al.*, 2003). Within the cell, long double-stranded RNA is cleaved into short 21–25 nucleotide siRNA by a ribonuclease known as Dicer. The siRNA subsequently assemble with protein components into an RNA-induced silencing complex (RISC). An ATP-dependent unwinding of the siRNA activates the RISC, which in turn binds to the complementary transcript by base pairing between the siRNA antisense strand and the mRNA. The bound mRNA is cleaved and results in gene silencing (McManus and Sharp, 2002). In addition to gene silencing, siRNAs also play diverse biological functions such as antiviral defense, transposon silencing, gene regulation, centromeric silencing and genomic rearrangement. **MicroRNA** (miRNA) is a small single-stranded (approximately 21–23 nt) RNA, whose function is to regulate gene expression by binding to the 3′–untranslated region (UTR) of target RNA to suppress translation or binding to the complementary sequence and thereby destroying the target transcript (Ambros, 2003).

4.6 RNA FOLDS AND STRUCTURE MOTIFS

4.6.1 RNA folds

Ribonucleic acids form regular and stable helical secondary structures from canonical base pairing. This continuous double-stranded RNA (dsRNA) presents a challenge for helix packing. The major groove, which displays the diverse base functional groups for tertiary interactions, is deep and narrow and thus difficult to access. The minor groove is broad and shallow without the chemical diversity required for interactions. Notwithstanding, RNAs adopt precise conformations in which the A-form helices pack together to produce unique structures capable of specific ligand recognition and catalysis. RNA-containing structures extracted from PDB and NDB are available at RNABase (http://www.rnabase.org/). There are several strategies that RNA might use to achieve a stable and compactly folded structure (Strobel and Doudna, 1997):

Helical stacking: Helical stacking is one strategy for packing RNA helices into a tertiary structure. The secondary structure of tRNA consists of four short helices that radiate from the center in a cloverleaf-like shape. In its three-dimensional structure, two pairs of helices coaxially stack and perpendicularly align to yield the L-shaped tertiary structure. Coaxial stacking of helices is observed in ribozymes leading to extensive tertiary interactions between helical subdomains.

Hydrogen bonding via 2′-Hydroxyls: The 2′-OH is the signature chemical group that distinguishes RNA from DNA. The 2′-OH groups of RNA lie on the outer edge of the minor groove where they can serve as either hydrogen bond donors or acceptors in tertiary interactions. Helix packing mediated by 2′-OH is often observed in the crystal structures of dsRNA. Each hydrogen bond involves pairs of 2′-OH where a hydroxyl on one helix serves as a hydrogen bond donor to a hydroxyl acceptor of the other helix.

Non-canonical base pairs: The helices in RNA are generally short (10 bp or less) and usually include unpaired nucleotides and non-canonical bps (bps other than G—C and A—U). Non-canonical bps present chemical groups in either the major or minor groove of the helix that can be used for unique tertiary interactions.

Metal ions: All RNAs require divalent cations for folding and catalysis because the negatively charged phosphate backbone is neutralized in order to facilitate the close packing of RNA helices. Metal ions are deeply embedded in RNA and are extensively coordinated or hydrogen bonded. These ions bind to specific sites within the RNA where they provide an ionic scaffold for tertiary structure formation.

Psudoknot constraints: The pseudoknot framework places loops and helices in close juxtaposition, which may promote formation of non-canonical structures. A pseudoknot can use the stability of Watson–Crick helices to pay for the unfavorable electrostatic free energy of bringing several strands close together. Nestled pseudoknots have been observed in the catalytic site of some ribozymes and mRNA (Glick and Draper, 1994). PseudoBase at http://wwwbio.leidenuniv.nl/~Batenburg/PKB.html is the database of RNA pseudoknots.

4.6.2 Structure motifs of RNA

RNA motifs refer to the small, finite and naturally occurring structural elements of RNAs that are present abundantly as non-canonical bps in hairpins/loops (loop motifs) and sequences involved in tertiary interactions (tertiary motifs). The most important secondary structural element in RNA is the A-form double helix (A-helix). However, motifs, which differ from A-helix, are discrete and their conformations are altered if residues are added or deleted from them. Some recognizable motifs found in RNA are considered (Moore, 1999).

4.6.2.1 Terminal loop motifs.

A terminal loop is any sequence where RNA folds back on itself so that a stem can form and highly abundant in terminal loops. They are:

U-Turn: The U-turn is common at the apices of the anticodon loops of tRNAs and an invariant feature of their Tψ loops. The consensus sequence for U-turn is unpaired UNRN (N for any bases and R for any purines). In Tψ loops, U is replaced by pseudouridine. The turns which are stabilized by hydrogen bonds between imino proton of U(1) and a phosphate oxygen of R(3), and between 2′-OH of U(1) and N_7 of R(3), introduce an abrupt 180° change in the backbone direction of larger loops.

Tetraloops: Three classes of 4-nucleotide terminal loops, UNCG, GNRA and CUYG (Y for any pyrimidines) referred to as tetraloops, are present in rRNAs. UNCG and GNRA tetraloops are similar in that non-canonical pairing between bases 1 and 4 reduces the distance between the backbone of the 5′- and 3′-sides of the stem, making the remaining gap easily bridged by bases 2 and 3. The 1–4 base pairing is an unusual syn-anti (U · G) in UNCG loops while it is side-by-side (G · A via AN_6—N_3G and AN_7—N_2G) in GNRA tetraloops. In CUYG tetraloops, the C(1)—G(4) pairing is Watson–Crick type but U(2) reaches down into the minor groove of the stem and interact with the adjoining G—C bp to form the six nucleotide consensus sequence, GCUYGC.

4.6.2.2 Internal loop motifs.

An internal loop is a sequence of bases within an A-helix stem that cannot form canonical pairs with its opposite strands consisting of any distinctive structure larger than a single, non-stranded bp found in an internal loop. Some recognized motifs are:

Cross-strand purine stacks: The cross-strand (rather than same-strand) stacking formed by the six-member ring of an A or G in one strand stacks on the six-member of an A or G in the other (either A's stack on A's or G's stack on G's). The consensus sequence for cross-strand A-stack motifs is 5′ (G/C)(G/A)A · 3′ (C/G)A(U/A). The initial (G · C) pair is Watson–Crick type. The next pair is usually side-by-side (A · G) or (A · A via AN_6—N_3A) and the following pair can be either reversed Hoogsteen (A · U vis UN_3—N_7, UO_2—N6A) or side-by-side AA. Additional nonhelical sequence must follow before an internal loop containing a cross-strand A-stack can be terminated and A-helix resumes.

One of the two cross-stranded G stacks is 5′UG · 3′GU. Any Watson–Crick pair will serve as the continuation on both flanks of this sequence, which is easily accommodated in A-helix because the UG pairs formed are wobbling. The six-membered rings of the G's in the central wobble pairs stack on top of each other, and the U's have no stacking partners on their 3′ sides. The other cross-strand G-stack is 5'CGA · 3'GAG. The six-membered rings of the G's in the two (A · G)'s stacks as well as A's stack also, with 5-membered ring of one stacking on the six-membered ring of the other.

Bulged G motifs: Bulged G motifs linked cross-strand A stacks with the reverse-Hoogtsteen AU type to A-helix. The consensus sequence is 5′GGA(A/C)Y · 3′CAUGAU. All the bases in this motif are paired except G(4) in the 3′ strand, which reaches across the major groove of the structure so that its imino proton can hydrogen bond with the phosphate group that links G(2)—A(3) in the 5′ strand. The base juxtaposition immediately beyond the bulged G are (A · A via AN_6—N_1A, and AN_1—N_6A) and (C · A via AN_1—N_4C and CN_1—N_6A). The backbone of the 5′ strand has a smooth, continuous trajectory, similar to that of A-helix. The 3′ strand is over-wound with a distinct S-turn. One or two (Y · Y) appears to provide flexibility for the Watson–Crick pairing to resume.

A-Platforms: A-Platforms, like bulged G motifs, are generally found in asymmetric loops, i.e. internal loops where the number of bases in each strand differ. A pair of neighboring A's in the sequence of one strand, which forms the platform, arrange so that their faces are nearly coplanar and oriented, such that their N_3—N_6 vectors are roughly parallel. The backbone connecting them is almost perpendicular to the axis of the helix. Either a (G · U) wobble pair or a non-Watson–Crick (A · U) pair is found below the platform, the purine of which belongs to the platform strand. Mg^{2+} or K^+ bond between the platform A's and the supporting (G · U)/(A · U) contributes to the stability of the motif. A-Platforms and bulged G motifs are asymmetric. Both are compatible with A-helix on one side but not the other.

Bulge-helix-bulge motif: It is a seven-base internal loop motif that interrupts what would otherwise be a continuous A-helix, as seen in the spliceosome. The consensus sequence including the flanking canonical pairs is 5′ (G/A)CUCNWRR(G/A)–3′ (C/U)ARAGRGN(C/U) (W for either A or U). The four bases at the 5′ ends of both strands form an A-helix four-bp, and the three bases at the 3′ ends of both strands form three-base bulges. The central helix of the motif comprises about a third of a turn of A-helix and in the vicinity of the two bulges the backbone of the bulge strand makes an approximate right-angle turn toward its partner. The minor groove face of the central helix is extended by the bulges at both ends to form a flat surface that possesses approximately twofold symmetry. It is this surface that the twofold symmetric splicing endonucleases attack during splicing (Trotta *et al.*, 1998).

Metal binding motifs: The folding of an RNA molecule that brings groups capable of coordinating metal ions (e.g. exocyclic oxygen atoms of bases, phosphate oxygens and nitrogens of the pyrimidine/purine rings) into an appropriate geometrical relationship may form a potential metal binding site. The binding of metal ions may contribute to the chemical activities of folded RNAs. Most metal binding sites have some shared properties. First, they are often found in the major groove of RNA stems. Second, the bases most likely to be involved in metal binding are U and C. The exocyclic oxygen atom at the 4 position of U is a good ligand, though the corresponding position in C is poor. The major groove face of G includes two good ligands, N_7 and O_6, but the major groove face of A has only one, N_7. Third, stems that have two successive G's in the same strand have increased potential for outer shell coordination for polyvalent metal ions.

4.6.2.3 Tertiary motifs.

Ribonucleic acids form regular and stable helical secondary structures from canonical base pairing, while additional non-canonical and tertiary interactions can give RNA its molecular shape. Thus RNA can be viewed structurally as a framework of helical segments that are linked by various non-canonical and tertiary interactions to produce a molecule with a specific three-dimensional fold. The strands in RNA stem-loops interact mostly by Watson–Crick pairing to form tertiary structure. Two larger-scale structural elements contribute to RNA tertiary structure; the coaxial stack and pseudoknot. In the coaxial stack, the blunt, nonloop ends of stem-loops that are adjacent in an RNA sequence tend to stack on top of each other so that the base stacks in one helix continue into the next without interruption. Bulges often form at such junctions so that coaxial stacking can occur. In the pseudoknot (a canonical pairing between a loop and a single strand in the same molecule), which refers to as the folding topology rather than specific structure, the loop bases of a stem-loop form a short double helix by interacting with either upstream or downstream sequences. Not only the stem-loop in pseudoknots varies in stem length and loop size, but also the length of sequences that interact with their loops. Another element, i.e. specific ion association, may also contribute to the formation of RNA (especially rRNA) tertiary structures. At lease two structure motifs stabilizing the tertiary structures are recognized.

Ribose zippers: Ribose zippers are stabilized by a hydrogen bond network that is formed between two antiparallel strands of RNA that come into close proximity but are not base paired. In this network, the hydrogens of the 2′OH's of the 5′ nucleotides in both strands interact with the oxygens of the 2′OH's of the 3′ nucleotides on the other strand. Hydrogen bonds also form between the 2′OH of the 3′ nucleotide on both strands and acceptor groups of the bases at the 5′ nucleotides of the other strand. The backbone conformation of a ribose zipper has twofold symmetry.

Tetraloop-helix interactions: RNA tertiary structures can be stabilized by hydrogen bonding interactions between the NRA bases of a GNRA tetraloop and groups found in the minor groove of helical stems, if their sequences (tetraloops) are appropriate. In the hammerhead structure of ribozymes, the GAAA tetraloop interacts with a helical stem capped with a GNRA tetraloop that ends with two Watson–Crick (G·C) pairs. The G's belonging to these GC pairs and the A in the capping GNRA interacts with the three A's of the incoming GAAA. The incoming A(2) interacts with the A of the stem tetraloop almost symmetrically. The tetraloop–tetraloop shares a base quadruple involving a (G·C) and the side-by-side (A·G) of A(4)—G(1) in the tetraloop. It is stabilized by hydrogen bonds between N_1 of A(4) and the 2′OH as well as the exocyclic amino group of the G in the GC pair. Furthermore, the 2′OH of the ribose of A(4) acts as both donor and acceptor to form hydrogen bonds with both the O_2 and the 2′OH of the C in the GC pair.

4.7 ENERGETICS OF NUCLEIC ACID STRUCTURE

A hydrogen bond is a short, noncovalent, directional interaction between a covalently bond H atom (donor) that has some degree of positive charge and a negatively charged (or electronegative) acceptor atom. In the DNA double helix, the N and O atoms involved in hydrogen bonding are separated by 2.82–2.92 Å (10^{-10} m), i.e. 2.82–2.91 Å for A·T pair (Seeman *et al.*, 1976) and 2.84–2.92 Å for G·C pair (Rosenberg *et al.*,1976). Typically, hydrogen bonds are weak, having only 12.6–29.3 kJ/mol (covalent bonds have 335–418 kJ/mol). In DNA, the hydrogen bonds have 8.4–12.6 kJ/mol. This is weaker than most hydrogen bonds and is due to geometric constraints within the double helix. Stability is largely determined by electrostatic and hydrophobic interactions between parallel overlapping base planes, which generate an attractive force called base stacking (Table 4.8). Since the aromatic bases are planar, they can stack nicely on one another. Hydrophobic

TABLE 4.8 Base pair stacking energies

Dinucleotide base pairs	Stacking energies (kcal/mol/stacked pair)
(GC) · (GC)	−14.59
(AC) · (GT)	−10.51
(TC) · (GA)	−9.81
(CG) · (CG)	−9.69
(GG) · (CC)	−8.26
(AT) · (AT)	−6.57
(TG) · (CA)	−6.57
(AG) · (CT)	
(AA) · (TT)	−5.37
(TA) · (TA)	−3.82

Note: Data adapted from Ornstein *et al.* (1978).

interactions and van der Waals forces are involved in the stacking interaction, which is estimated to be 16.7–62.8 kJ/mol per nucleotide. Van der Waals interaction involves dipole-dipole interactions and London dispersion interactions (transient dipole interactions). Differences between the characteristics of base stacking and hydrogen bonding energies contribute to the heterogeneity of the DNA helix structure. The overall energy of hydrogen bonding depends predominantly on base composition. However, base stacking energies depend on the sequence of the DNA. Thermodynamic data for nucleic acids can be retrieved from NTDB at http://ntdb.chem.cuhk.edu.hk/ (Chiu *et al.*, 2003).

The melting temperature (T_m) of nucleic acids defines the thermal point at which 50% of the complementary molecule dissociate/melt away from their cognate strands or anneal/hybridize the two strands. The capacity to estimate melting temperatures is important because it helps the researcher select strand separation/hybridization temperatures, such as in polymerase chain reaction and microarray assays. A good approximation of melting temperature for oligonucleotides takes the expression (Wallance *et al.*, 1979):

$$T_m = 4(G + C) + 2(A + T)$$

For example, the melting temperature of an octadecanucleotide with the sequence AGTTACCTACAATCAGGC would be $T_m = 4\ (8) + 2\ (10) = 52°C$. This provides a good estimate for relatively short oligonucleotide (5–20 mer) hybridized in aqueous buffers containing low salt (<100 mM Na^+). However, predicted melting temperatures for any oligonucleotides can be calculated on-line (www.gensetoligos.com/Calculation/calculation.html) by submitting the nucleotide sequence. It provides, in addition to the melting temperature, information concerning oligonucleotide length, GC content and absorption coefficient at 260 nm.

The classic Watson–Crick base pairing scheme is only one of several. There are many other ways in which two bases can be held together by hydrogen bonding, such as reversed Watson–Crick, Hoogsteen, reversed Hoogsteen, Wobble and G (*anti*)-A (*syn*) base-pairing schemes. The bases are in chemical equilibrium between the two alternative tautomeric forms, keto and enol (for G and T) or amino and imino (for A and C). The equilibrium favors the keto and amino forms by a ratio of about 10^4 to 1. Tautomerization creates miss-pairs in the DNA. In addition to keto-enol/amino-imino tautomerizations, bases can exist in ionized forms that also change their hydrogen bonding properties. When ionization occurs, a number of non-Watson–Crick bps can arise. Adenine is prone to protonation by low pH, which can lead to the formation of an $A^+ \cdot C$ bp leading to the wobble pairing scheme. Cytosine is also prone to protonation. Protonation of C can lead to a $C^+ \cdot G$ bp, which requires a Hoogsteen pairing scheme.

4.8 NUCLEIC ACID APPLICATION

Two examples of novel applications of DNA/RNA will be considered here, i.e. the use of DNA/RNA as aptamers and the use of DNA in nanoengineering. Aptamers are molecules selected to have high affinity binding to a preselected target (Brody and Gold, 2000). The plasticity of single-stranded nucleic acid polymers combined with the affinity of complementary bases to base pair provides an enormous potential for variation in three-dimensional (tertiary) structure. These properties have been used to identify nucleic acids that bind tightly to specific ligands such as proteins, peptides or organic molecules. DNA and RNA are ideal aptamers because powerful methods exist to work with complex mixtures of species and to purify from these mixtures just the molecules with high affinity for a target from which to characterize the common features of these classes of molecules. Novel nucleic acid catalysts can also be selected for the variants that make or break covalent bonds. Aptamers are isolated from complex libraries of synthetic oligo- or polynucleotides by an iterative process of adsorption, recovery and amplification, known as systematic evolution of ligands by exponential enrichment (SELEX) to be described in subsection 15.4.2. Aptamers which act in the same way as antibodies by folding into 3D structures based on their sequences and bind to various target molecules with high affinity without displaying no immunogenicity, make them attractive therapeutic agents (Nimjee *et al.*, 2005). Aptamer database can be accessed at http://aptamer.icmb.utexas.edu/.

The goal in nanoengineering is to make machines, motors, transducers, and tools at the molecular level (Fortina *et al.*, 2005). The potential advantages of using DNA as construction material in nanosciences (Niemeyer, 2000) are due to several factors:

- The enormous specificity of the A—T and G—C hydrogen bonding allows the convenient programming of designed DNA receptors.
- The great versatility to synthesize DNAs of desired sequence by automated methods and to amplify any sequence from microscopic to macroscopic quantities by means of PCR.
- The ability to make complex two- and three-dimensional arrays using Holliday junctions (Holliday, 1964) and the formation of structures of accurate length by taking advantage of the great stiffness of duplex DNA.
- The great mechanical rigidity of short DNA duplex enables DNA to behave effectively as a rigid rod spacer between two tethered functional molecular components on both ends.
- DNA displays a relatively high physicochemical stability.
- The availability of highly specific enzymes allows for the processing of DNA material with atomic precision and accuracy on the molecular level.
- The potential for anchoring proteins along the DNA at many points creates arrays with more complex properties.

No other polymeric material offers these advantages, which are ideal for molecular construction in the range of nanometers to micrometers. The disadvantage of using DNA and proteins for nanoengineering is that these structures are mechanically easier to deform than typical atomic or molecular solids. They usually require an aqueous environment, and have relatively limited electrical or mechanical properties. There are two potential ways to circumvent the potential disadvantages of DNA in nanoengineering. One is to use DNA (and proteins) to direct the assembly of other types of molecules with the desired properties needed to make engines or transducers. A second is to use the DNA as a resist: to cast a solid surface or volume around it, remove the DNA and then use the resulting cavity as a framework for the placement of other molecules. In both applications, DNA is really conceived of as the ultimate molecular scaffold, with adjustable lengths and shapes.

4.9 REFERENCES

ABOUL-ELA, F., MURCHIE, A.I.H., NORMAN, D.G. *et al.* (1994) *Journal of Molecular Biology*, **243**, 458–71.

ADAMS, M.D., FIELDS, C. and VENTER, J.C. (1994) *Automated DNA Sequencing and Analysis*, Academic Press, San Diego, CA.

ALPHEY, L. (1997) *DNA Sequencing*, Springer-Verlag, New York.

AMBROS, V. (2003) *Cell*, **113**, 673–6.

ANSORGE, W., VOSS, H. and ZIMMERMANN, J. (1997) *DNA Sequencing Strategies*, John Wiley & Sons, New York.

Arnott, S. and Selsing, E. (1974) *Journal of Molecular Biology*, **88**, 509–21.

Beerman, T.A. and Lebowitz, J. (1973) *Journal of Molecular Biology*, **79**, 451–70.

Bhattacharyya, A. and Lilley, D.M.J. (1989) *Nucleic Acids Research*, **17**, 6821–40.

Blackburn, G.M. and Gait, M.J. (1997) *Nucleic Acids in Chemistry and Biology*, IRL Press, Oxford, UK.

Bloomfield, A., Crothers, D.M. and Tinoco, I. Jr. (2000) *Nucleic Acids: Structures, Properties and Functions*, University Science Books, Sausalito, CA.

Bowater, R., Aboul-ela, F. and Lilley, D.M.J. (1991) *Biochemistry*, **30**, 11495–506.

Breslauer, K.J., Frank, R., Bloccker, H. *et al.* (1986) *Proceedings of the National Academy of Sciences, USA*, **83**, 3746–50.

Brody, E.N. and Gold, L. (2000) *Journal of Biotechnology*, **74**, 5–13.

Brown, T.A. (1994) DNA Sequencing: A Practical Approach, IRL Press, Oxford, U.K.

Brumbaugh, J.A., Middendorf, L.R., Grone, D. *et al.* (1988) *Proceedings of the National Academy of Sciences, USA*, **85**, 5610–14.

Calladine, C.R. (1982) *Journal of Molecular Biology*, **161**, 343–52.

Chargaff, E. (1951) *Federal Proceedings*, **10**, 654–9.

Chiu, W.L., Abe, K., Sze, C.N. *et al.* (2003) *Nucleic Acids Research*, **31**, 483–5.

Cotton, R.G.H., Rodrigues, N.R. and Campbell, R.D. (1988) *Proceedings of the National Academy of Sciences, USA*, **85**, 4397–401.

Dam, E., Pleij, K. and Draper, D. (1992) *Biochemistry*, **31**, 11665–76.

Dickman, S. and Wang, J.C. (1985) *Journal of Molecular Biology*, **186**, 1–11.

Dickerson, R.E., Drew, H.R., Conner, B.N. *et al.* (1982) *Science*, **216**, 475–85.

Dickerson, R.E., Bansal, M., Calladine, C.R. *et al.* (1989) *Journal of Molecular Biology*, **205**, 787–79

Eguchi, Y., Itoh, T. and Tomizawa, J. (1991) *Annual Reviews in Biochemistry*, **60**, 631–52.

Felsenfeld, G. and Miles, H.T. (1967) *Annual Reviews in Biochemistry*, **26**, 407–68.

Fortina, P., Kricka, L.J., Surrey, S. *et al.* (2005) *Trends in Biotechnology*, **4**, 168–173.

Galazka, G., Palecek, E., Wells, R.D. *et al.* (1986) *Journal of Biological Chemistry*, **261**, 7093–8.

Glazer, A.N. and Mathies, R.A. (1997) *Current Opinions on Biotechnology*, **8**, 94–102.

Glick, T.C. and Draper, D.E. (1994) *Journal of Molecular Biology*, **241**, 246–62.

Hagerman, P.J. (1985) *Biochemistry*, **24**, 7033–7.

Hanvey, J.C., Shimizu, M. and Wells, R.D. (1988) *Proceedings of the National Academy of Sciences, USA*, **85**, 6292–6.

Harborth, J., Elbashir, S.M., Vandenburgh, K. *et al.* (2003) *Antisense Nucleic Acid Drug Developments,* **13**, 83–106.

Herr, W. (1985) *Proceedings of the National Academy of Sciences, USA*, **82**, 8009–13.

Holliday, R. (1964) *Genetic Research*, **5**, 282–304.

Hoogsteen, K. (1963) *Acta Crystallography*, **16**, 907–16.

Howell, *et al.* (1966) *Biochemistry*, **35**, 15373–82.

Htun, H. and Dahlberg, J.E. (1989) *Science*, **243**, 1571–6.

Hutvagner, G. and Zamore, P.D. (2002) *Science*, **297**, 2056–60.

Johnston, B.H. (1988) *Science*, **241**, 1800–4.

Kim, J., Cheong, C. and Moore, P.B. (1991) *Nature*, **351**, 331–2.

Klysik, J., Zacharias, W., Galazka, G. *et al.* (1988) *Nucleic Acids Research*, **16**, 6915–33.

Lafer, E.M., Möller, A., Nordheim, A. *et al.* (1981) *Proceedings of the National Academy of Sciences, USA*, **78**, 3546–50.

Larsen, B.S. and McEwen, C.N. (eds) (1998) *Mass Spectrometry of Biological Materials*, 2nd edn, Marcel Dekker, New York.

Lauglan, G., Murchie, A.I.H., Norman, D.G. *et al.* (1994) *Science*, **265**, 520–4.

Letai, A.G., Palladino, M.A., Fromm, E. *et al.* (1988) *Biochemistry*, **27**, 9108–12.

Lilley, D.M.J. (1985) *Nucleic Acids Research*, **13**, 1443–65.

Lilley, D.M. (1989) *Chemical Society Reviews*, **18**, 53–83.

Lyamichev, V.I., Mirkin, S.M., and Frank-Kamenetskii, M.D. (1986) *Journal of Biomolecular Structural Dynamics*, **3**, 667–9.

Maxam, A.M. and Gilbert, W. (1977) *Proceedings of the National Academy of Sciences, USA*, **74**, 560–4.

Maxam, A.M. and Gilbert, W. (1980) *Methods in Enzymology*, **65**, 499–560.

Maxwell, E.S. and Fournier, M.J. (1995) *Annual Reviews in Biochemistry*, **64**, 897–934.

McClellan, J.A., Palecek, E. and Lilley, D.M.J. (1986) *Nucleic Acids Research*, **14**, 9291–309.

McClellan, J.A. and Lilley, D.M. (1991) *Journal of Molecular Biology*, **219**, 145–9.

McKeon, C., Schmidt, A. and de Crombrugghe, B. (1984) *Journal of Biological Chemistry*, **259**, 6636–40.

McManus, M.T. and Sharp, P.A. (2002) *Nature Reviews of Genetics*, **3**, 737–47.

Mirkin, S.M., Lyamichev, V.I., Drushlyak, K.N. *et al.* (1987) *Nature*, **330**, 495–7.

Moore, P.B. (1999) *Annual Reviews in Biochemistry*, **67**, 287–300.

Murchie, A.I.H. and Lilley, D.M.J. (1987) *Nucleic Acids Research*, **15**, 9641–54.

Murchie, A.I.H., Bowater, R., Aboul-ela, F. *et al.* (1992) *Biochimera Biophysica Acta*, **1131**, 1–15.

Nadal, M., Mirambeau, G., Forterre, P. *et al.* (1986) *Nature*, **321**, 256–8.

Neidle, S. (2002) *Nucleic Acid Structure and Recognition*, Oxford University Press, Oxford, UK.

Niemeyer, C.M. (2000) *Current Opinions in Chemical Biology*, **4**, 609–18.

Nimjee, S.M., Rusconi, C.P. and Sullenger, B.A. (2005) *Annual Reviews in Medicine*, **56**, 555–83.

Nordheim, A., Pardue, M.L., Lafer, E.M. *et al.* (1981) *Nature*, **294**, 417–22.

Nordheim, A. and Rich, A. (1983) *Proceedings of the National Academy of Sciences, USA*, **80**, 1821–5.

Orr-Weaver, T.L., Szostak, J.W. and Rothstein, R.J. (1981) *Proceedings of the National Academy of Sciences, USA*, **78**, 6354–8.

Rice, J.A. and Crothers, D.M. (1989) *Biochemistry*, **28**, 4512–6.

Rich, A. (1977) *Acc. Chem. Resear.*, **10**, 388–96.

Rich, A. Nordhiem, A. and Wang, A.H.-J. (1984) *Annual Reviews in Biochemistry*, **53**, 791–846.

Saenger, W. (1984) *Principles of Nucleic Acid Structure*, Springer-Verlag, New York.

Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *Proceedings of the National Academy of Sciences, USA*, **74**, 5463–4.

Schimmel, P. (1987) *Annu. Rev. Biochem.*, **56**, 126–158.

Schimmel, P. (1989) *Cell*, **58**, 9–12.

Scovell, W.M. (1986) *Journal of Chemistry Education*, **63**, 562–5.

Sen, D. and Gilbert, W. (1990) *Nature*, **344**, 410–14.

Singleton, C.K., Kilpatrick, M.W. and Wells, R.D. (1984) *Journal of Biological Chemistry*, **259**, 1963–7.

Smith, A.J.H. (1980) *Methods in Enzymology*, **65**, 560–80.

Smith, L.M., Sander, J.Z., Kaiser, R.J. *et al.* (1986) *Nature*, **321**, 674–9.

Sprinzl, M., Dank, N., Nock, S. and Schon, A. (1991) *Nucleic Acid Resear.*, **19**, 2127–71.

Steitz, T.A. (1990) *Quarterly Reviews in Biophysics*, **23**, 205–80.

Storz, G., Altuvia, S. and Wassarman, K.M. (2005) *Annual Reviews in Biochemistry*, **74**, 199–217.

Streisinger, G., Okada, Y., Emrich, J. *et al.* (1966) *Cold Spring Harbor Symposium on Quant. Biology*, **31**, 77–84.

Strobel, S.A. and Doudna, J.A. (1997) *Trends in Biochemical Sciences*, **22**, 262–6.

Sullivan, K.M. and Lilley, D.M. (1987) *Journal of Molecular Biology*, **193**, 397–404.

Sundquist, W.I. and Heaphy, S. (1993) *Proceedings of the National Academy of Sciences, USA*, **90**, 3393–7.

Sundquist, W.I. and Klug, A. (1989) *Nature*, **342**, 825–9.

Thomas, B. and Akoulitchev, A.V. (2006) *Trends in Biochemical Sciences*, **31**, 173–81.

Trotta, C., Mario, F., Arn, E. *et al.* (1998) *Science*, **280**, 279–84.

Wang, A.H.J., Quigley, G.J., Kolpak, F.J. *et al.* (1979) *Nature*, **282**, 680–6.

Wang, A.H.-J., Fujii, S., van Boom, J.H. *et al.* (1982) *Nature*, **299**, 601–4.

Wang, Y., de los Santos, C., Gao, X. *et al.* (1991) *Journal of Molecular Biology*, **222**, 819–32.

Watson, J.D. and Crick, F.H.C. (1953) *Nature*, **171**, 737–8.

Wilkins, M.H.F., Stokes, A.R. and Wilson, H.R. (1953) *Nature*, **171**, 738–40.

Williams, W.L. and Müller, U.R. (1987) *Journal of Molecular Biology*, **196**, 743–55.

Wittig, B., Wolf, S., Dorbic, T. *et al.* (1992) *EMBO Journal*, **11**, 4653–63.

Wu, H.-M. and Crothers, D.M. (1984) *Nature*, **308**, 509–13.

Wu, J. and McLuckey, S.A. (2004) *International Journal of Mass Spectrometry*, **237**, 197–241.

Zakian, V.A. (1989) *Annual Review of Genetics*, **23**, 579–604.

Zucker, M. (1989) *Science*, **244**, 48–52.

World Wide Webs cited

Aptamer database:	http://aptamer.icmb.utexas.edu/
DNA Data Bank of Japan:	http://www.ddbj.nig.ac.jp
EBI of EMBL:	http://www.ebi.ac.uk
GenBank of the NCBI, NIH:	http://www.ncbi.nlm.nih.gov/Genbank
Non-canonical base pair database:	http://prion.bchs.uh.edu/bp_type/bp_structure.html
Nucleic Acid Database (NDB):	http://ndbserver.rutgers.edu
Oligonucleotide calculation:	http://www.genseloligos.com/Calculation/calculation.html
PseudoBase:	http://wwwbio.leidenuniv.nl/~Batenburg/PKB.html
Rfam:	http://www.sanger.ac.uk/Software/Rfam.
RNABase:	http://www.rnabase.org/
RNADB:	http://jsm-research.imb.uq.au/rnadb
RNA modification database:	http://medlib.med.utah.edu/RNAmods
SCOR:	http://scor.lbl.gov/
Thermodynamic data (NTDB):	http://ntdb.chem.cuhk.edu.hk/

CHAPTER 5

BIOMACROMOLECULAR STRUCTURE: PROTEINS

5.1 ARCHITECTURE OF PROTEIN MOLECULES

5.1.1 Introduction

Proteins are the agents of biological function. Virtually every cellular activity is dependent on one or more proteins (Creighton, 1993; Fersht, 1999; Whitford, 2005). Therefore, they can be classified functionally as catalytic proteins (enzymes), regulatory proteins, transport proteins, storage proteins, structural proteins, contractile proteins, scaffold proteins, protective proteins and others. Furthermore, based on their shape and solubility, they can be assigned to one of three global classes, namely fibrous, globular or membrane. Fibrous proteins often serve as structural constituents of cells such as collagen, fibroin and α-keratin. They have regular linear structures and are insoluble in water or dilute salt solutions. In contrast, globular proteins are roughly spherical in shape and are usually soluble in aqueous solutions. Most of soluble proteins of the cell such as cytosolic enzymes and regulatory proteins are globular proteins. Some regulatory proteins and transport proteins are found in association with cellular membranes, which generally have helical cores and are insoluble in aqueous solutions but can be solubilized with detergents.

The architecture of protein molecules is complex and can be described according to structural organization as primary structure (amino acid sequence), secondary structure (regular structures such as helical, pleated sheet, and coil structures), tertiary structure (fold in three-dimensional space), quaternary structure (subunit structure) and quinternary structure (biomacromolecular complexes). Usually the overall three-dimensional (3D) architecture of a protein molecule is termed as its conformation, which refers to its secondary and tertiary structures. Between these two structures, motifs (supersecondary structures) refer to the packing of adjacent secondary structures into distinct structural elements and domains refer to identifiable 3D structural units that may correspond to functional units. The structures of most proteins with more than 200 amino acid residues appear to consist of two or more domains.

5.1.2 Representation of protein structures

Amino acid residues of polypeptide chains are generally represented by the three-letter codes or the one-letter codes. The sequences are written from the left for the N-terminus to the right for the C-terminus, e.g. human lysozyme:

MKALIVLGLVLLSVTVQGKVFERCELARTLKRLGMDGYRGISLANWM
CLAKWESGYNTRA

Biomacromolecules, by C. Stan Tsai
Copyright © 2007 John Wiley & Sons, Inc.

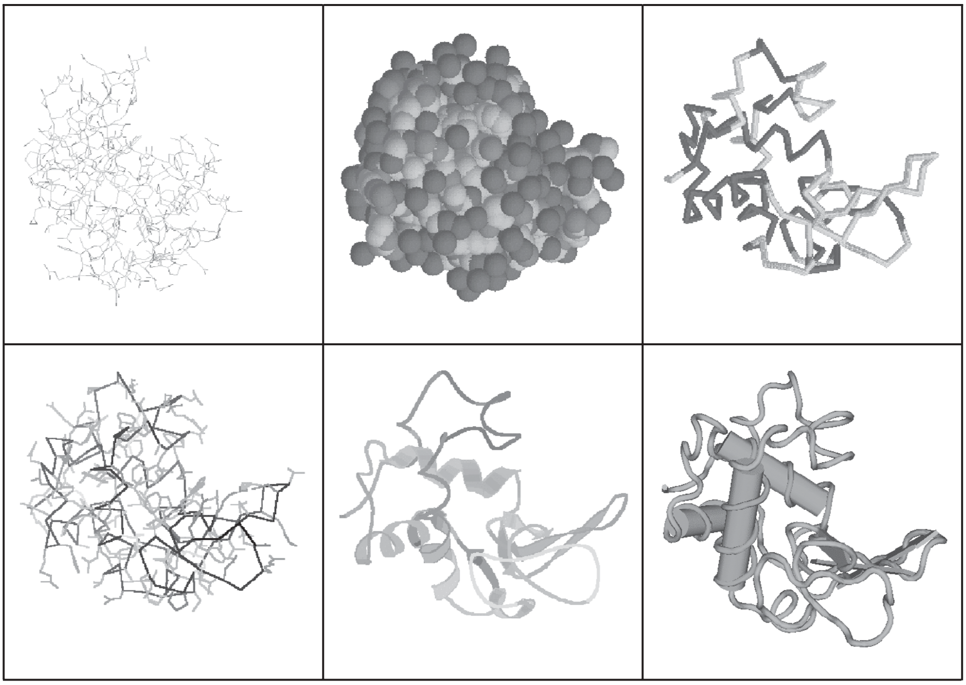


Figure 5.1 Common representations of the three-dimensional structures of proteins
Three-dimensional graphics of hen's egg-white lysozyme as visualized with RasMol (first row, 1LYZ.pdb) are displayed in wireframe (residue type), spacefill (atom type) and backbone (structure type). The C_α mainchain with side chain residues is displayed with KineMage (second raw left). The secondary structure representations (α helices and β-sheets) are visualized with KineMage (1LYZ.kin) in ribbons and arrows (second raw center), and Cn3D (1LYZ.cn3) in cylinders and arrows (second raw right).

TNYNAGDRSTDYGIFQINSRYWCNDGKTPGAVNACHLSCSALLQDNIA DAVACAKRVVRD

PQGIRAWVAWRNRCQNRDVRQYVQGCGV

Amino acid sequences of proteins can be retrieved from the Protein Information Resources (PIR)-International Protein Sequence Database (PSD) at http://pir.georgetown.edu, Swiss-Prot of Swiss Institute of Bioinformatics at http://expasy.hcuge.ch/sprot/ and UniProt consortium at http://www.uniprot.org. All natural proteins possess specifically folded 3D structures and a large number of these protein structures have been determined by atomic resolution. The Protein Data Bank (PDB) at http://www.rcsb.org/pdb/ is the global archive of structural data of proteins and other biomacromolecules (Kouranov, A. *et al.*, 2006). Computer graphics provides the best and most comprehensive tool for depicting the three-dimensional structures of biomacromolecules (Figure 5.1). Three popular freeware programs for molecular graphics are Cn3D (http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.html), KineMage (http://kinemage.biochem.duke.edu/) and RasMol (http://www.umass.edu/microbio/rasmol/).

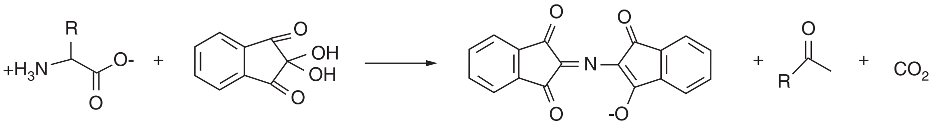
5.2 PRIMARY STRUCTURE OF PROTEINS: CHEMICAL AND ENZYMATIC SEQUENCE ANALYSIS

The primary structure of protein (i.e. amino acid sequence) dictates high-level protein structures (secondary, tertiary and quaternary structures) and determines their chemical and biological properties. The general strategy for the amino acid sequence determination of a protein involves the following basic steps (Needlman, 1975; Findley and Geisow, 1989; Smith, 2002):

1. Separation and purification of individual peptide chains from proteins containing multiple chains.
2. Cleavage of intrachain disulfide linkages by either oxidation (e.g. HCOOOH) or reduction (e.g. $NaBH_4$). This step precedes step 1 if the disulfides are interchain linkages.
3. Determination of amino acid composition.
4. Identification of the N-terminal and C-terminal residues.
5. Cleavage of polypeptide chain into smaller manageable fragments, specifically and reproducibly.
6. Purification of the peptide fragments and analysis of their amino acid compositions.
7. Determination of amino acid sequences of the peptide fragments from the N-terminus and/or the C-terminus.
8. Repeat steps 5–7, using a different cleavage procedure to generate a different and therefore overlapping set of peptide fragments.
9. Construction of the overall amino acid sequence of the protein from the sequences in overlapping fragments.
10. Localization of the disulfide linkage of the polypeptide chain.

5.2.1 Amino acid composition

The analysis of amino acid composition (Barrett, 1985) can be performed by complete hydrolysis of a protein sample, followed by quantitative analysis of the liberated amino acids. Besides hydrolysis with 6 M hydrochloric acid at 120°C for 12 h or with dilute alkali (2–4 M NaOH) at 100°C for 4–8 h, mixtures of peptidases can also be used for complete peptide hydrolysis. Pronase consisting of a mixture of relatively unspecific peptidases from *Streptomyces griseus* is often used for enzymatic hydrolysis. However, the amount of peptidase used should not be more than ~1% by weight of the polypeptide to be hydrolyzed. Otherwise, self-degrading by-products might contaminate the final digest. The enzymatic procedure is mostly used for the determination of Trp, Asn and Gln. The original automated equipment separates amino acids by ion-exchange chromatography, and employs post-column derivatization with ninhydrin to give the blue color of Ruhemann's violet.



The modern automated amino acid analyzer based on HPLC (Hancock, 1984) can perform complete analysis in less than 60 m with a sensitivity as low as 1 pmol of each amino acid. Various derivatization protocols for high performance (high-pressure) chromatography (HPLC) use fluorescent and UV-absorbing tags, such as 1-dimethylamino phthalene-sulfonyl (dansyl), 9-fluorenylmethoxycarbonyl (Fmoc), *iso*-indolyl and Edman's reagent.

5.2.2 Peptide cleavage, separation and analysis

To facilitate sequence analysis, the polypeptide chain is cleaved into fragments of manageable size reproducibly with enzymatic and/or chemical methods (Table 5.1).

The choice of protease depends on the sample and method. Trypsin is perhaps the most commonly chosen protease because so much is known about its behavior. In general, it is a good idea to maintain a trypsin-to-sample ratio to 1 : 50 to 1 : 100. Trypsin itself and its autodigested peptides could contaminate the peptide map, so running a protease control is essential. It is also useful to have the sequence of trypsin (and other proteases) available to prevent accidentally assigning a protease sequence to a sample. A good strategy is to reduce the sample-protease ratio (<100 : 1) and time of digestion (perhaps 1 to 4 h). Peptide analysis is best performed by reversed phase HPLC on narrow bore (2.1 mm i.d.), microbore (1 mm i.d.) or capillary column (<1 mm i.d.) packed with C-4, C-8 or C-18 packing materials. The highest sensitivity is observed for packed capillary columns using methods described by Simpson *et al*. (1989) or Davis and Lee (1992) (Figure 5.2).

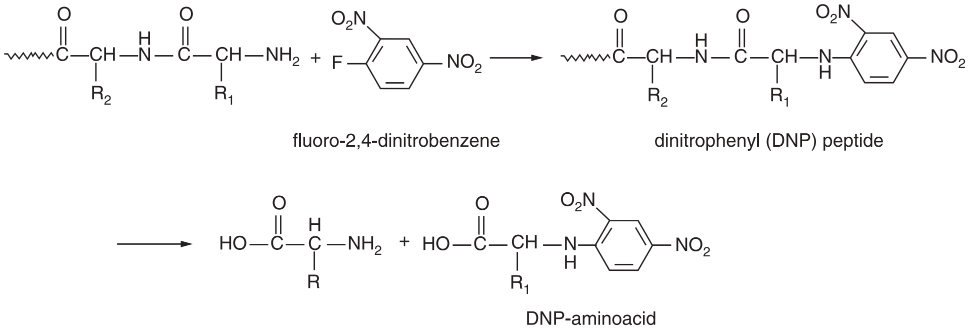
5.2.3 Terminal and sequence determination

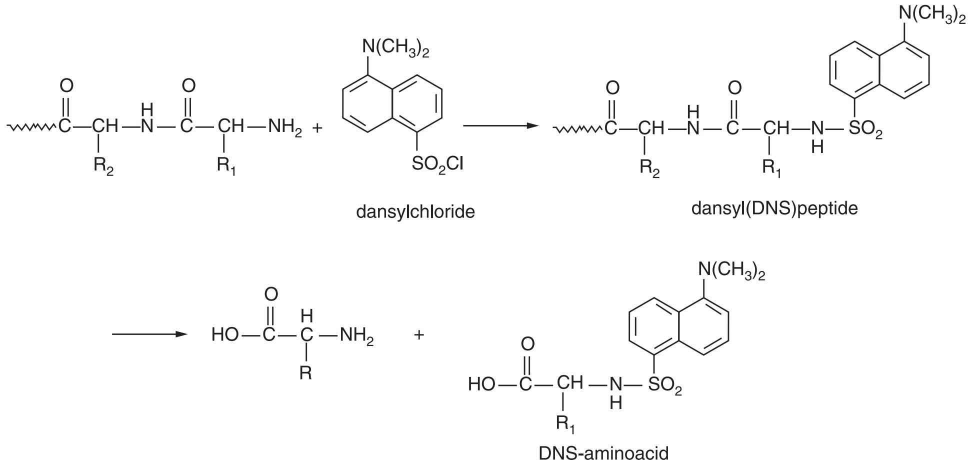
The common methods for determining N-terminal residue employ 1-fluoro-2,4-dinitrobenzene (Sanger's reagent) and 1-dimethylamino phthalene-5-sulfonyl (dansyl) chloride. The derivatized peptide is hydrolyzed, and the labeled N-terminal residue is identified by its yellow color as dinitrophenyl (DNP) amino acid or by fluorescence as dansyl (DNS) amino acid.

TABLE 5.1 Specificity of cleavage agents

	Cleavage agent	Site (N→C) cleaved
Chemical:	Hydroxalamine	Asn-Gly
	Mild acid, pH 2.5	Asp-Pro
	Cyanogen bromide	Met-X
	N-Chlorosuccinimide	Trp-X
Enzymatic:	Chymotrypsin (pancreas)	Aromatic, e.g. Phe-X, Trp-X, Tyr-X
	Trypsin (pancreas)	Arg-X, Lys-X
	Bromelin, elastase (pancreas)	Small neural, e.g. Ala-X, Gly-X
	Pepsin (gastric mucosa)	Leu-X, Phe-X, hydrophobic
	Subtilisin	Aliphatic-X, aromatic-X
	Staphylococcus protease	Glu-X, Asp-X
	Prolylendopeptidase	Pro-X
	Papain	Y-Arg, Y-Lys, Y-Gly
	Thermolysin	Y-Aliphatic, Y-Phe
	Asp-N-protease	Y-Asp
	Aminopeptidase	N-terminal
	Carboxypeptidase	C-terminal

Note: X and Y represent any amino acids.


fluoro-2,4-dinitrobenzene
dinitrophenyl (DNP) peptide
DNP-aminoacid


dansylchloride
dansyl(DNS)peptide
DNS-aminoacid

Procedures for determining the C-terminal residue of proteins involve hydrazenolysis and reduction with hydrides (directly or after esterification), though they are not widely used. The C-terminal residue is identified as either the amino acid or its alcohol after hydrolysis. In the hydrazenolysis, the peptidic hydrazide can be removed by extracting with benzaldehyde, which forms hydrazone with hydrazide.

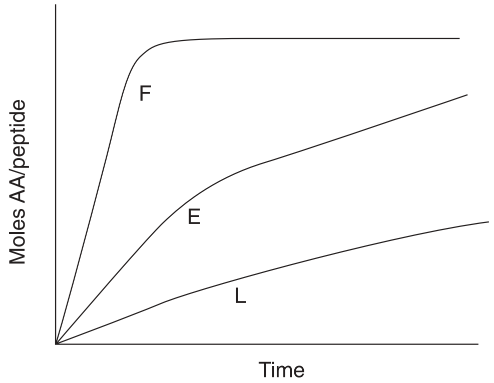
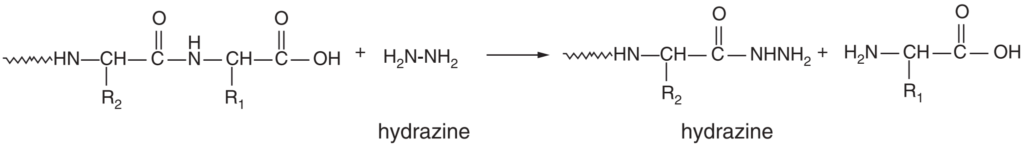
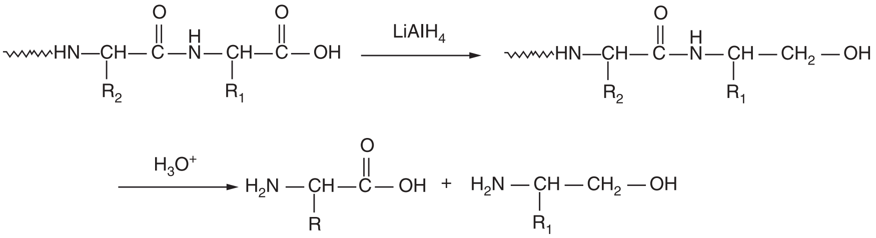


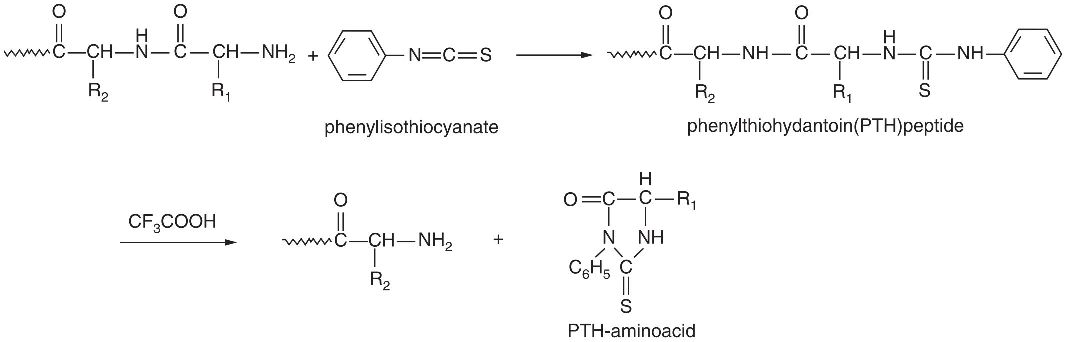
Figure 5.2 Liberation of C-terminal amino acids from peptide by carboxypeptidase
An example of the C-Terminal determination uses carboxypeptidase which liberates amino acids sequentially from peptide–LeuAspPhe$_{COOH}$.



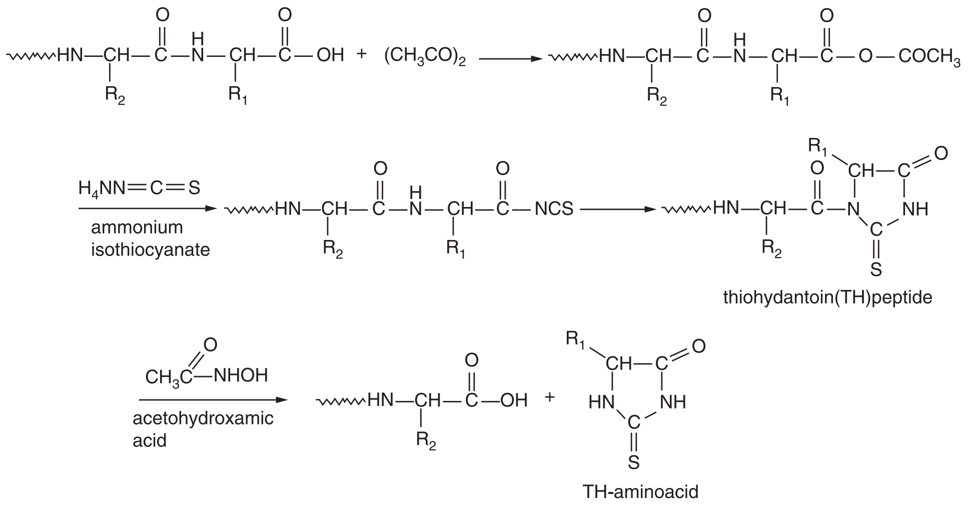


The peptide sequence determination can be preceded via stepwise (cyclic) degradation/identification of N-terminal residues or C-terminal residues. This is accomplished by either chemical or enzymatic methods.

5.2.3.1 Chemical methods. In chemical methods, free N-terminal or C-terminal amino acids are coupled/labeled with a specific reagent, which aids in the subsequent degradative process of the derivatized (labeled) terminal residue. The cleavage of the labeled terminal amino acid generates the shorten peptide with a new terminal residue for the next degradation cycle. The most successful method for the determination of amino acid sequences is Edman degradation that releases UV absorbing phenylthiohydantoin (PTH) amino acids stepwise from the N-terminus, which are readily identified (Edman and Begg, 1967).



An analogous Stark degradation releases the C-terminal residues successively as thiohydratoin (TH) amino acids (Bailey and Shively, 1990).



In the first automated sequence-determination instrument applying Edman degradation (Edman and Begg, 1967), the protein sample was contained in a spinning cup that produced a thin film, over which the reagents and solvents passed. The technique of solid-phase Edman degradation (Hunkapiller and Hood, 1980) was followed by an introduction of the gas-phase sequencer (Hewick *et al.*, 1981). The instrument featured a cartridge assembly that contained a low-volume (50 μL) reaction chamber designed to hold a Polybrene-coated glass fiber filter. To improve the overall sensitivity, protein was electroblotted on to either Polybrene-coated glass fiber sheets (Aebersold *et al.*, 1986) or a polyvinylidene difluoride membrane (Matsudaira, 1987). The development of HPLC gave analysts a method with good sensitivity and the ability to resolve all PTH-amino acid derivatives in a single analysis (Hunkapiller and Hood, 1978). Edman sequanators are completely automated high-sensitivity instrument systems that can routinely detect and identify as little as 0.2 to 1 pmol of amino acid in a given cycle of operation.

5.2.3.2 Enzymatic methods. The use of aminopeptidases and carboxypeptidases that removes amino acids sequentially from the N-terminus and C-terminus of a peptide chain respectively, though technically simple is limited in practical application. In chemical methods, peptide chains in the sample all undergo first cycle of degradation before the second cycle is initiated. This is not the case in enzymatic degradation, which is its main limitation. The order of the amino acid residues is not determined in a stepwise fashion, but rather from the rate at which the amino acids appear in the digest, i.e. an amino acid appearing faster presumably precedes the slower ones in the sequence. Furthermore, the specificity of the enzyme may prevent or retard the release of particular terminal amino acids (e.g. C-terminal proline and arginine residues are not susceptible to carboxypeptidase A) and therefore limits its application.

The most popular method is automated Edman chemistry (run on a sequencer), a method which removes one amino acid at a time from the N-terminus, resulting in the sequential liberation of phenylthiohydantoin (PTH) amino acids, which are identified by on-line HPLC analysis. Edman sequanators are completely automated high-sensitivity instrument systems that can routinely detect and identify as little as 0.2 pmol to 1 pmol of amino acid in a given cycle and carry out more than 20 cycles with 1 to 5 pmol of protein. Most peptides of length 3–30 amino acids can be sequenced completely. Although amino acid sequences of proteins, especially sequences of insoluble or rare proteins, are derived from gene sequences, protein sequencing techniques are still required to determine the partial sequences that are then used to design oligonucleotides that are complementary to parts of the genes that encode the proteins.

5.2.4 Peptide ladder sequencing

Efficient sequence analysis requires fragments of the protein to be analyzed, since stepwise degradation is limited to 40–80 residues. Therefore longer polypeptide chains must be cleaved into fragments suitable for stepwise analysis by either enzymatic or chemical methods. Peptide ladder sequencing combines ladder-generating chemistry and mass spectrometry (Chait *et al.*, 1993). N-terminal ladder sequencing requires a modified Edman procedure in which the peptide is incompletely degraded to continuously yield a mixture of one-amino acid-shortened peptides in every step. Such a peptide ladder is formed when Edman degradation with phenyl isothiocyanate (PITC) is performed in the presence of 5% phenyl isocyanate (PIC) as the terminating reagent. The phenylthiocarbamoyl peptide is cleaved in the presence of a strong acid to give the 5(4H)-thiazoline, while the phenylcarbamoyl peptide, formed from PIC, is stable under these conditions. During the next cycles, the content of phenylcarbamoyl peptides increases to a statistical mixture, thereby forming the peptide ladder.

Analysis of the mixture using matrix-assisted desorption ionization mass spectrometry (section 5.3) allows for direct sequence determination from the successive mass differences of the peptide ladder. The application of the volatile trifluoroethyl isothiocyanate results in a significant optimization of this procedure and allows for peptide sequencing at the femtomole level (Bartlet-Jones *et al.*, 1994). C-terminal ladder sequencing uses ammonium thiocyanate in acetic anhydride coupled with mass spectrometric analysis of truncated peptides (Thiede *et al.*, 1997). Matrix-assisted desorption ionization instruments with delayed extraction (Brown and Lennon, 1995) allow for the discrimination of all amino acids, except Leu and Ile.

5.3 PRIMARY STRUCTURE OF PROTEINS: SEQUENCE ANALYSIS BY TANDEM MASS SPECTROMETRY

5.3.1 An application of mass spectrometry (MS) in protein chemistry

In the electron spray ionization (ESI) MS (Whitehouse *et al.*, 1985), an acidic aqueous peptide/protein solution (e.g. 1 to 20 μL/m of water-acetonitrile containing 0.05 to 0.1% acetic acid or TFA; the purpose of the acid is to encourage protonation of the sample) is sprayed through a small-diameter needle. The use of acidic conditions in ESI to positively charged droplets tends to protonate all available basic sites. The maximum charge state usually corresponds to the total number of basic groups in a peptide. As a result, multiply protonated peptide ions are observed whenever a lysine, arginine or histidine residue is present in a peptide because one proton is associated with the N-terminal amine and additional protons associate with each additional basic residue. Doubly charged peptides tend to predominate in tryptic digests of proteins because of the proteolytic specificity of trypsin that cleaves amide bonds at the C-terminal lysine or arginine residue. Ionization takes place by protonation of these two sites. More highly charged trypic peptides nearly always contain internal histidine, lysine-proline bonds, arginine-proline bonds or missed cleavage sites. As a result, the maximum charge state of a peptide in a tryptic digest can provide some information about its structure. Furthermore, the higher charge states focus according to m/z, and they are seen at lower masses on the mass scale. This is an obvious advantage for larger peptides and proteins, which may otherwise be out of the mass range of analyzer. Since all of the peaks for a given sample are related by one charge unit between peaks, the mass spectrum may be simplified by deconvolution, that is, combining all of the charge states for a single molecule into a single peak. Most data systems perform this task automatically. Negative ion spectra may also be recorded, but requires special conditions. The sample is infused in the presence of a volatile basic counter ion, such as aqueous ammonia. In addition, a flow of air or oxygen is required to suppress coronal discharge. Negative ion spectra may be especially useful for glycopeptide and sulfated or phosphorylated peptides. ESI sources are found on quadrupole and magnetic sector mass spectrometers.

Nanospray (Wilm *et al.*, 1996) and microspray (Emmett and Caprioli, 1994) are additional spray-ionization techniques used in the MS analysis of protein digests. The ionization technique operates on principles similar to ESI with the significant reduction of flow rates and needle diameters used for the spray. Whereas electrospray operates with microliter-per-minute flow rates through needle with inside diameters in the 100 μm to 200 μm range, nanospray operates with nanoliter-per-minute flow rates through 5 μm to 10 μm (i.d.) needles. As a result, a greater proportion of analyte is desorbed from the droplets and is transmitted from the spray needle to the entrance aperture of the mass spectrometer, resulting in the detectable signals at the attomole amounts of peptides (Table 5.2). This sensitivity enhancement extends the application of nanospray ionization to sequencing of proteins in electrophoretic gels down to the silver stain-detectable level, which is equivalent to as little as 10 to 100 fmol of protein in the gel. The advantage of microspray is its compatibility with in-line, capillary-column liquid chromatography, because these columns are amenable to elution with these flow conditions.

For matrix-assisted laser desorption ionization (MALDI), peptide/protein are dissolved in a solution of a UV-absorbing compound, referred to as the matrix, and placed on a probe or stage for the mass spectrometer. The sample (0.1 to 10 pmoles in 1 μL) is mixed with an UV-absorbing compound referred to as a matrix, such as nicotinic acid, sinapinc acid (3,5-dimethoxy-4-hydroxycinnamic acid), α-cyano-4-hydroxycinnamic acid or 2,5-dihydroxybenzoic acid and allowed to crystallized on a sample stage. The sample is ionized by laser desorption, usually with a nitrogen laser (337 nm). The UV absorbing matrix vaporizes and ionizes, carrying the peptide/protein sample along with it. Ionization occurs by protonation in the acidic environments produced by the acidity of most matrix compounds and by the addition of dilute acid to the sample. If laser power is properly controlled, the major signal observed is for the protonated molecular ion. MALDI is ideally suited for time-of-flight (TOF) mass spectrometers, since the laser firings can be used as timed events (i.e. can start the clock for mass measurements). For TOF m/z analysis, an ion is given a fixed amount of kinetic energy by acceleration in an electric field that is

TABLE 5.2 Experimental conditions for electrospray, micospray and nanospray ionization

	Electrospray	Microspray	Nanospray
Typical flow rate, μL/min	2	0.2	0.02
Needle size (i.d.), μm	100	75	5
Approx. limit of detection	10 femtomole	1 femtomole	50 attomole

generated by the application of a high voltage, typically +20 kV to +30 kV. Following acceleration, the ion enters a field-free region where it travels at a velocity that is inversely proportional to its m/z. Because of this inverse relationship, ions with low m/z travel more rapidly than ions with high m/z. The time required for the ion to travel the length of the field-free region is measured and used to calculate the velocity and ultimately the m/z of the ion according to $v = ((2 * V * z)/m)^{1/2}$, where v = velocity in m/sec, V = accelerating voltage in J, m = mass in kg, and z = charge. TOF mass analysis requires that the set of ions being studied be introduced into the analyzer in a pulse. As a result, TOF mass analysis is ideally suited to ionization technique like MALDI that produce ions in short, well-defined pulses. TOF-MS have very high sensitivity due to their high ion transmission rates and it is not unusual to detect fmole amounts of peptide or protein in MALDI-TOF instruments.

The mass spectra obtained by ESI and MALDI differ, depending on the resolution of the instrument and the presence of multiply charged species. The mass for a given peptide may be calculated in two ways. The so-called exact or monoisotopic mass is the mass calculated assuming all the carbon is C-12, all nitrogen is N-14 and all of the hydrogen is H-1. In reality, 1.1% of the carbon is C-13, 0.37% of the nitrogen is N-15 and 0.015% of the hydrogen is H-2 (negligible for most MS). A low-resolution instrument such as TOF or quadrupole instruments will not resolve the peaks corresponding to the isotopic abundance, and thus average mass calculations are appropriate. For high-resolution instruments, such as magnetic sector or FT-MS, it is usually possible to resolve the monoisotopic peak, and with tandem instruments, select it for further analysis. ESI measurements usually give a mass accuracy of 0.01 to 0.05%. Mass measurements of proteins by MALDI-TOF are usually accurate to ±0.05 to 1.0%.

5.3.2 Application of tandem mass spectrometry (MS–MS) in protein sequence analysis

Tandem mass spectrometer uses the separation of ions according to their m/z as a preparative tool to isolate an ion with a specific m/z for further analysis (McLafferty, 1983). This further analysis is carried out by fragmenting the mass-selected ion and by determining the m/z of the fragment ions in a second stage of mass analysis. The term tandem mass spectrometry (MS–MS) reflects the fact that two stages of mass analysis are used in a single experiment. The result is that specific ions in a complex mixture can be selectively studied in an experiment that gives structural information about that ion. Ions formed by any ionization process are classified according to their stability over the course of the mass measurement process as either stable, unstable or metastable. Stable ions remain intact and do not fragment. Conversely, unstable ions have sufficient internal energy to fragment immediately to form stable product ion. Metastable ions have an intermediate amount of excess internal energy and fragment after moving from the ion source into the mass analyzer. The ionic products of the fragmentation reaction are referred to as product ions or fragment ions. In MS experiments, information about the structure of an analyte is derived from the observation of product ions. Because the majority of ions formed by ESI and MALDI are stable, little structural information is present in the mass spectra that are produced by these ionization techniques. Furthermore, because fragmentation does not occur, the MS–MS would simply re-measure the m/z of the mass-selected ion a second time. What is required is a method to energize a stable ion after it has been mass-selected and to induce informative fragmentation reactions. The process used most often is collisional activation, in which a mass-selected ion is transmitted into a high-pressure region of the tandem mass spectrometer where it undergoes a number of collisions with gas molecules.

A portion of the kinetic energy of the ion is converted into internal energy in the ion to make the ion unstable and drive fragmentation reactions that occur prior to leaving the collision cell. This process is called either collisionally activated dissociation (CAD) or collisionally induced dissociation (CID). The fragment ions produced are m/z analyzed in the second stage of mass analysis. The precursor-ion selection effectively isolates a single peptide ion from the mixture and removes any contribution of the other peptides to the sequence analysis step (Kinter and Sherman, 2000). Furthermore, this isolation step is accomplished on the microsecond time scale, directly in-line with the subsequent structural analysis steps, promoting the speed and sensitivity of the analysis.

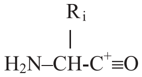
Tandem mass spectrometers are classified as either tandem-in-space or tandem-in-time. Instruments in the tandem-in-space including tandem quadrupole, quadrupole-time-of-flight, reflection-time-of-fight, tandem sector and sector-quadrupole, have more than one mass analyzer and each mass analyzer performs separately to accomplish the different stages of the experiment. Instruments that perform tandem-in-time, including ion trap and Fourier transform, have only one mass analyzer and, by necessity, this mass analyzer must be capable of trapping ions. MS–MS experiments are carried out in three scan modes; product ion scan, precursor ion scan and neutral loss scan. The product ion scan mode is the experiment in which fragmentation data are recorded to determine the amino acid sequence of a peptide. The role of precursor ion scans or neutral loss scans is generally to aid in the selection of specific ions for subsequent product ion scans.

In MS sequencing, the information that describes the amino acid sequence of a peptide is contained in a product ion spectrum. This product ion spectrum is obtained in an MS–MS experiment by using CID of a protonated or multiply protonated peptide ion. Routine complete interpretation of product ion spectra to deduce the entire sequence of a peptide ion has been mainly replaced by database search programs that search large databases of spectra derived from the protein and translated gene sequences. The search identifies peptide sequences in the databases that are consistent with the spectrum.

Understanding the structure of protonated peptides and their fragmentation pathways plays a key role in interpreting product ion spectra. Peptide bonds may be broken in the mass spectrometer by either high or low energy collision processes (collision induced dissociation, CID). In magnetic sector instruments, the ions are accelerated at high potential (10 to 20 kV) and thus fragment extensively when expose to a collision gas. Fragments occur on both sides of the peptide bonds. In order to get complete sequence information; CID should produce fragments on both sides of the peptide bond for every amino acid in the peptide (Figure 5.3).

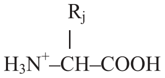
The nomenclature used to describe the different product ions defines two sets of ions that are named based on the peptide terminus retained in the ion (Biemann, 1990). In this nomenclature, the a-, b- and c-ions all contain the N-terminus of the peptide, while the x-, y- and z-ions all contain the C-terminus. The major N-terminus containing ion series is the b-ion series and the major C-terminus containing ion series is the y-ion series. They are the most useful and the most common sequence ions. The b ions arise from peptide bond cleavage and transfer of the positive charge to the C-terminal side of an amino acid (the acyl group carries the positive charge), resulting in a fragment ion series starting from the N-terminus. y ions arise from peptide bond cleavage and transfer of the positive charge to the N-terminal side of an amino acid (the amino group carries the positive charge), resulting in a fragment ion series starting from the C-terminus (BiEmann, 1990).

For example,



$$\mathrm{H_2N{-}\overset{\overset{R_i}{|}}{C}H{-}C^{+}{\equiv}O}$$

b_i ion



$$\mathrm{H_3N^{+}{-}\overset{\overset{R_j}{|}}{C}H{-}COOH}$$

y_j ion

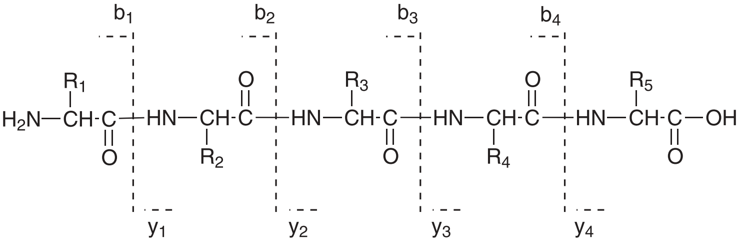


Figure 5.3 MS fragmentations of peptide linkage

TABLE 5.3 Useful information in the fragmentation chemistry of peptides

Amino acid	Residue mass (Da)	Immonium ion (m/z)	Equivalent amino acid pair	Neutral ion loss
Glycine (G)	57.02	30		–
Alanine (A)	71.04	44		–
Serine (S)	87.03	60		18
Proline (P)	97.05	70		–
Valine (V)	99.07	72		–
Threonine (T)	101.05	74		18
Cystine (C)	103.01	76		34
Leucine (L)	113.08	86		–
Isoleucine (I)	113.08	86		–
Asparagine (N)	114.04	87	GG	17
Aspartic acid (D)	115.03	88		18
Glutamine (Q)	128.06	101	GA	17
Lysine (K)	128.09	101	GA	17
Glutamic acid (E)	129.04	102		18
Methinine (M)	131.04	104		48
Histidine (H)	137.06	110		–
Methionine sulfoxide (Mo)	147.04	120		64
Phenylalanine (F)	147.07	120		–
Arginine (R)	156.10	129	GV	17
Tyrosine (Y)	163.06	136		–
Tryptophan (W)	186.08	159	GE, AD, SV	–

In theory, a complete set of b and y ions contains all of the necessary sequence information, but in practice, complete sets of either ion series are rarely observed. Two pairs of amino acids have the same mass, namely Leu and Ile, and Lys and Gln. In order to distinguish Leu and Ile, we must rely on fragments within the side chain, fragments, which are only seen in high-energy CID. Lys and Gln having the same mass can be distinguished by analyzing the sample before and after acetylation (that specifically modifies Lys), resulting in a mass shift for Lys residues and the N-terminal amino acid of 42 (CH_3CO group) mass units (m.u.). To distinguish N-terminal fragments from C-terminal fragments, the protein is acetylated, the N-terminal amino acid fragment will increase in mass by 42. However, if the C-terminus is Lys, the C-terminal fragment will also increase by 42 m.u. Another approach is to treat the sample with methanolic HCl, a procedure that converts peptide carboxyl groups to their methyl esters (an increase in 14 m.u. per carboxyl group). This procedure not only allows us to count carboxyl groups, but may also help to locate the C-terminal series.

Table 5.3 lists the 20 common amino acids and the monoisotopic residue mass given to two decimal places. These residue mass values are used when interpreting product ion spectra, although it is easiest to use the simple nominal mass when working through a spectrum interpretation. The difference between nominal and monoisotopic mass becomes significant, as a series of values are totaled and produce what is essentially a rounding error. We should always base the specific amino acid assignments on the residue masses calculated from the measured m/z of the ions in the spectrum and use the calculated nominal masses as an aide to track the interpretation process. In the table, methionine sulfoxide (oxidized methionine) is included because it is commonly observed. It will be noted that, in general, the residue masses of amino acids vary considerably, from 57 Da for glycine to 186 Da for tryptophan. however, there is some overlap among the residue masses of several amino acids. First Leu and Ile have identical residue masses and under low-energy CID conditions these two amino acids cannot be distinguished. For the purposes of spectrum interpretation, these two amino acids are treated as one and given the one-letter abbreviation X to denote that either amino acid is possible at that position. Second, Gln and Lys both have nominal residue masses of 128 Da, and phenylalanine and oxidized methionine both have nominal residue masses of 147 Da. These amino acid pairs can often be differentiated and assigned by using other MS data. For example, Gln and Lys can be distinguished by an acetylation reaction. It is also common to assume that, in tryptic peptides, Lys should be found only at the C-terminus; and if an internal Lys IS present, then the peptide ion will add another proton and exist as a triply charged ion. The difference between phenylalanine and oxidized methionine can be less clear. If a single phenylalanine or oxidized methionine residue mass is observed, we may be able to distinguish between the two based on the ion resulting from the loss of the small neutral molecule, $HSOCH_3$ (64 Da) from oxidized methionine or from the immonium ion for phenylalaine at m/z 120. Alternatively, CNBr, which specifically cleaves proteins at methionine, may be used to produce CNBr-cleaved peptides for the analysis. The other residue masses of the amino acids all differ by at least 1 Da and should be distinguishable when unit mass resolution or better is used. The table also lists the masses of the immonium ions ($H_2N{=}CHR^+$). Although these ions are informative about amino acid content, they provide no information about the position of that amino acid in the sequence. Also lack of an immonium ion is not usually conclusive. For example, the presence of m/z 110 indicates a high probability of a histidine somewhere in the peptide chain but the absence of m/z 110 does not preclude the presence of histidine in a peptide.

Although the b_1 ion is rarely seen in a product ion spectrum, the b_2 ion is almost always present, provided an appropriate m/z range can be used. Furthermore, because of the facile loss of CO, the b_2-ion is generally observed paired with the a_2-ion (resulting from b_2-ion by a loss of CO) and is recognizable by the 28 Da (m.u.) difference between the two ions. Whereas observation of the b_2-ion provides a limited set of possible combinations for the two N-terminal amino acids, precise assignment of identity and order of these two amino acids depends on observing the y_{n-1}-ion, which should be remembered when considering combinations of two amino acids. In a limited number of cases, the combined residue mass of two amino acids may equal the residue mass of a single amino acid. For example, the combined residue mass of GE, AD and SV are all equal to the 186 Da residue mass of a tryptophan residue.

The table also summarizes a set of information that may be useful when interpreting product ion spectra, mass differences produced by the loss of a small neutral molecule. These losses are observed for both the b- and y-ion series if that ion contains a particular amino acid. As a result, observation of these losses can be used as a general indicator of amino acid content of the peptide or product ion. The most common loses are water (18 Da) and ammonia (17 Da). Loss of water typically occurs from the alcohol- (serine and threonine) or carboxylic acid- (asparatate and glutamate) containing amino acids but may also be lost from the C-terminus of y-ions. Ammonia losses are generally due to the presence of the amide-(asparagine and glutamine) or amine-(lysine and arginine) containing amino acids but also occur from the N-terminus. Observing these types of ions is most useful when following a given ion series because the losses will be seen only as long as the amino acid producing that neutral loss is retained in the product ions. For example, b-ions of serine or threonine-containing peptides will generally be accompanied by ions 18 Da less than the particular b-ion. A change in the pattern of neutral loss within b-ions containing Ser/Thr often indicates that one of these amino acids is the next amino acid in that series. Another common neutral loss is 64 Da ($HSOCH_3$) from oxidized methionine-containing peptides or product ions. Table 5.3 also gives some useful information on peptide fragmentation.

The strategy for interpretation of the product ion spectra of tryptic digests is:

1. *Inspect the low-mass region for immonium ions*. The first step in the interpretation is to inspect the low-mass region of the spectrum, observing the presence of any immonium ions ($H_2N{=}CHR^+$) and the amino acid composition that they indicate.
2. *Inspect the low-mass region for the b_2-ion*. In the next step, the low-mass region of the spectrum is inspected to identify the b_2-ion, which is generally recognizable by the b_2-ion/a_2-ion pair separated by 28 Da. The a-ions result from the loss of CO from b-ions and are seen as b–28 Da ions. In the product ion spectra, only the b_2-ion routinely generates a detectable amount of the a_2-ion. After identifying the b_2-ion, its m/z is used to calculate the m/z of the corresponding y_{n-2}-ion, and the high-mass region of the spectrum is inspected to identify this ion.
3. *Inspect the low-mass region for the y_1-ion*. To assign the C-terminal amino acid, the low-mass region of the spectrum is inspected to identify the y_1-ion at either m/z 147, for C-terminal lysine peptides, or m/z 175 for C-terminal arginine peptides. The m/z of the y_1-ion is then used to calculate the m/z of the b_{n-1}-ion, and the high-mass region of the product ion spectrum is inspected to identify that ion.
4. *Inspect the high-mass region to identify the y_{n-1}-ion*. Attempt to assign the N-terminal amino acids from combinations indicated by the b_2-ion. The high-mass region of the spectrum is scrutinized to identify the y_{n-1}-ion, if present. The list of possible amino acid combinations derived from the b_2-ion limits the possible residue masses to consider. If an ion is identified, the m/z of that ion is used to calculate the residue masses of the first two amino acids and to assign those peptides.
5. *Extend the y-ion series toward lower m/z*. Extend the y-ion series backwards (toward lower m/z) from the y_{n-2}-ion via working with the residue masses of amino acids (Table 5.3). As a y-ion is identified, calculate the m/z of the corresponding b-ion and identify that ion in the spectrum. Work toward extending the y-ion series from the y_{n-2}-ion to the y_1-ion.
6. *Extend the b-ion series toward higher m/z*. If progress extending the y-ion series falters, use the residue masses of amino acids to extend the b-ion series from the last identified b-ion. As any b-ions are identified, use the m/z of that ion to calculate the m/z of the corresponding y-ion and identify that ion in the spectrum.
7. *Calculate the mass of the peptides*. When the interpretation of the spectrum is complete, calculate the mass of the proposed peptide sequence and check its agreement with the measured mass.
8. *Reconcile the amino acid content with spectrum data*. Check that the amino acid content agrees with the immonium ions observed. Also consider the charge state of

the peptide in terms of the presence of histidine, and internal lysine or arginine residues.

9. *Attempt to identify all ions in the spectrum.* Proceed to identify the other ions in the spectrum based on the proposed peptide sequence and pay attention to the ions from the loss of H_2O, NH_3 and $HSOCH_3$ as well as doubly charged ions and any ions due to internal cleavage.

In positive-ion operating conditions, ESI produces peptide ions that enter the mass spectrometer with protons attached to all of the strongly basic sites in the peptide. These sites include the N-terminal amine and the side group of any lysine, arginine or histidine residues. In the gas-phase, a proton associated with the Lys, Arg and His is strongly attached and remains associated or fixed at that site even on collisional activation. In contrast, a proton on the N-terminus may move by internal solvation to any of the amide linkages. As a result, a given protonated peptide formed by ESI is best viewed as a heterogeneous population of peptide ions, in which different sub-populations of ions have the same amino acid sequence but with a proton associated with each amide linkage. After mass selection, this population of peptides is then accelerated into multiple collisions with gas molecules in the collision cell of tandem mass spectrometers. The site of the protonated amide bonds directs the subsequent fragmentation reactions. Thus the variety of protonation sites directs the fragmentation reactions to occur at each of the different amide bonds. The manner by which the structure of protonated peptides directs the fragmentation is known as the mobile proton hypothesis (Dongre *et al*., 1996). The interpretation of a product ion spectrum begins with the assumption that the monoisotopic molecular weight of the peptide is in singly protonated form, $(M + H^+)$, and the most abundant charge state of the peptide from the MS is known.

Since the sensitivity of MS methods is similar to that of Edman chemistry, both methods will continue to be used. Because MS methods cannot determine the N- or C-terminal sequences in intact proteins, there will be a continued need for the Edman sequencer. However, MALDI- and ESI–MS methods can give accurate molecular masses for intact proteins, which when compared to their predicted sequences and databases (Eng *et al*., 1994), can verify a given structure, giving confidence to the N- and C-terminal sequence predicted from the peptide maps. MS-MS provides the fast and sensitive approach to sequence determination of peptides/proteins.

5.4 CONFORMATIONAL MAP

Various non-superimposable 3D arrangements of atoms in space, that are interconvertible without breaking covalent bonds, are described as conformations. Many bonds in biomacromolecules can rotate, resulting in many conformations. Rotations about bonds are described as torsion or dihedral angles, which are usually taken to lie in the range of −180° to +180°. The two dihedral angles associated with the polymer along the main chain, ϕ and Ψ are quite useful in defining the path of the polymeric chain atoms. In theory, a chain can adopt an essentially infinite variety of backbone conformations, each corresponding to a unique set of values for ϕ and Ψ. In fact, the allowed conformations are limited by steric restrictions, i.e. allowed conformations permit contact distances between non-neighboring atoms. Databases for conformational angles in peptides and proteins are available at PepConfDB (http://www.peptidome.org/products/list.html) and CADB (http://cluster.physics.iisc.emet.in/cadb/) respectively. The allowable values of ϕ and Ψ are usually represented on the ϕ versus Ψ plot, known as the conformational map or Ramachandran plot (diagram) as illustrated in Figure 5.4.

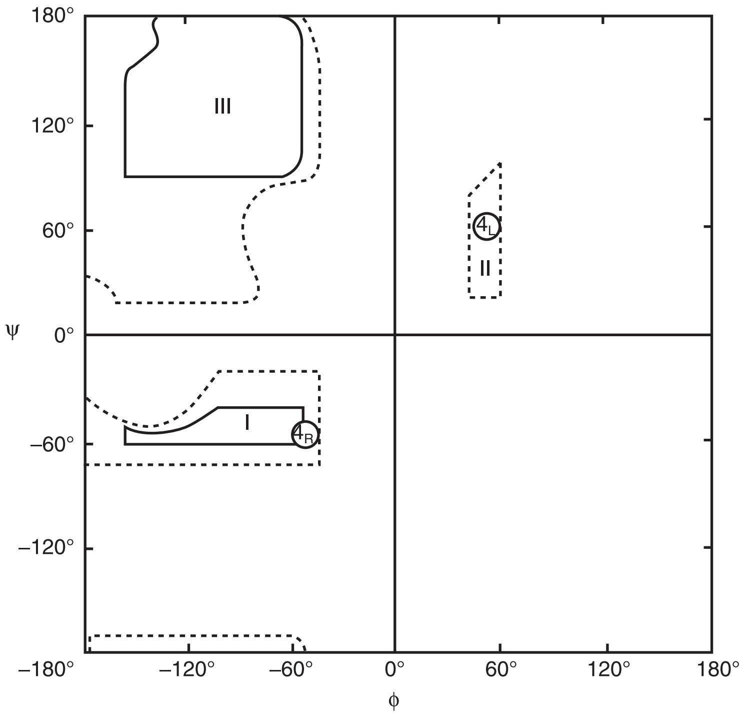


Figure 5.4 Conformational map for protein
The coordinates are ϕ (phi), ψ (psi) angles defined as: 1. Keeping two peptide units (planes) fully stretched and coplanar giving $\phi = 0°$ and $\psi = 0°$, 2. From N-terminus, holding the first unit rigidly and rotating an angle ϕ about the bond N—C_α in a clockwise manner looking from N toward C_α, 3. Holding the second amino acid residue firmly and rotating an angle ψ about the bond C_α—C_O in a clockwise manner looking from C_α toward C_O, 4. Generating a conformation of the two-linked peptide units corresponding to the pair of dihedral angles, ϕ and ψ, 5. Repeating the procedure along the peptide chain starting from N-terminus toward C-terminus. The conformation of the polypeptide-chain backbone is specified by enumerating the value for each ϕ_i and ψ_i for ith amino acid residue. The enclosed regions (I–III) represent the allowed conformations which are limited by steric restrictions (contact distances between atoms). The area connecting regions I and III represents region IV which is allowed if the $C_\alpha$angle, τ is slightly extended. Most ϕ, ψ pairs fall within or close to the allowed regions (points in nonallowed regions tends to be Gly).

The peptide bond has several important properties. It is close to being planar with a barrier of 79.5 kJ/mole for interconversion between the *cis* and *trans* conformations. The isomerization rate at 313 K is ~$0.15\,s^{-1}$ for model compounds, and activation free energy of ~79 kJ/mole. The equilibrium favors the *trans* conformation by a factor of 10^3, except in the case where the nitrogen is part of the five-membered ring of Pro, in which case it is favored only by a factor of 4. A second important property of the peptide bond is its dipole moment of ~3.5 Debye units. This moment, the associated hydrogen bonding properties of the carbonyl oxygen and peptide nitrogen atoms, and the consequent impact on their solvation drive the formation of secondary structure and have other fundamental consequences for the structure of proteins and the function of enzyme active sites. Finally, the stereochemistry of the peptide unit generates much of the hierarchic organization in protein structure.

Proteins are polymers comprised exclusively of L-amino acids. When the alpha carbon (C^α) is viewed from the hydrogen atom, rotation from C=O (C′) to R to N is clockwise for L-amino acids, giving rise to the mnemonic 'CORN' (i.e. CO→R→N). The two dihedral angles associated with the polymer, ϕ and Ψ are useful in defining the path of the main chain atoms (Ramachandran *et al.*, 1963). For polypeptide chains, the values of ϕ and Ψ associated with each of amino acid residue are defined by:

1. Keeping two peptide units (planes) associated with the amino acid residue fully stretched and coplanar (this defines $\phi = 0°$ and $\Psi = 0°$).
2. From N-terminus, holding the first unit (residue) rigidly and rotating an angle ϕ about the bond $N—C^\alpha$ in a clockwise manner, looking from N toward C^α (this defines ϕ value).
3. Holding the second unit (residue) firmly and rotating an angle Ψ about the bond $C^\alpha—C'$ in a clockwise manner, looking from C^α toward C' (this defines Ψ value).
4. Generating a conformation of the two linked peptide units (amino acid residues) corresponding to the pair of dihedral angles ϕ and Ψ (both ϕ and Ψ are defined).
5. Repeating the procedure along the peptide chain from the N-terminus toward the C-terminus; the conformation of the polypeptide chain as a whole is specified by enumerating the values of each ϕ_i and Ψ_i for *i*th amino acid residue.

With the bond angle for each amino acid residue, $\tau[NC_i^\alpha C'] = 110°$, three allowed regions (I, II and III) are constructed. An area between regions I and III may form the region IV if the angle τ ($C^\alpha—C^O—N$) is slightly increased (from 110° to 115°). X-ray crystallographic results indicate that this region is allowed if τ is increased to 115°. Amino acid residues falling within the regions of the conformation map are likely to adopt specific conformations:

Region	Conformation/structure
I	right-handed α (3.6_{13}), 3_{10} and π (4.4_{16}) helices
II	left-handed α (3.6_{13}), 3_{10} and π (4.4_{16}) helices
III	β sheets (parallel and antiparallel), triple helix
IV	miscellaneous

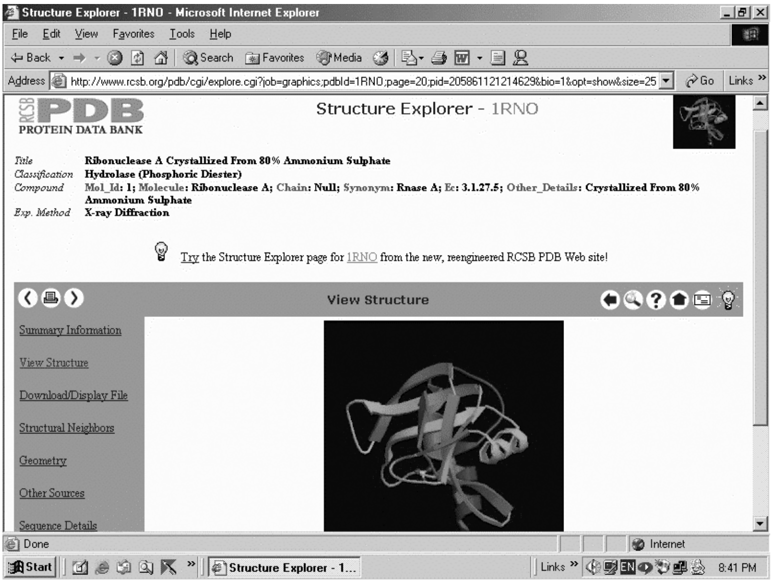
Many of the amino acid residues in the non-allowed region correspond to glycine.

An alternative convention useful in analyzing geometry of the backbone chain entails the description of pseudobonds between successive C^α atoms (Oldfield and Hubbard, 1994), because they form dihedral angle involving four consecutive C^α atoms. This formalism captures concisely more characteristics of the space-curve traced by the main chain than do the Ramachandran angles. Thus it has become a standard representation in studies of protein folding (Skolnick *et al.*, 1997).

5.5 SECONDARY STRUCTURES AND MOTIFS OF PROTEINS

Proteins exhibit regularities at several levels, which can be identified only by the fact that they recur approximately, but frequently, in different contexts. Much insight into the structures of proteins is therefore statistical in nature, coming from the comparison of large numbers of different proteins. The atomic coordinates for all known protein structures are deposited at the Protein Data Bank and can be retrieved from http://www.rcsb.org/pdb/ (Figure 5.5).

The secondary structures imply the hierarchy by providing repeating sets of interactions between functional groups along the polypeptide backbone chain that creates, in turn, irregularly shaped surfaces of projecting amino acid side chains. The secondary structures in proteins arise from repeating patterns of similar peptide dihedral angles (ϕ and Ψ) for successive residues (Table 5.4). There is a thermodynamic tendency for successive


Structure Explorer - 1RNO - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address
Go
Links
RCSB PDB
PROTEIN DATA BANK
View Structure
Summary Information
View Structure
Download/Display File
Structural Neighbors
Geometry
Other Sources
Sequence Details
Done
Internet
Start
Structure Explorer - 1...
Links
8:41 PM

http://www.rcsb.org/pdb/cgi/explore.cgi?job=graphics;pdbId=1RNO;page=20;pid=205861121214629&bio=1&opt=show&size=25

Structure Explorer - 1RNO

Title **Ribonuclease A Crystallized From 80% Ammonium Sulphate**
Classification **Hydrolase (Phosphoric Diester)**
Compound **Mol_Id: 1; Molecule: Ribonuclease A; Chain: Null; Synonym: Rnase A; Ec: 3.1.27.5; Other_Details: Crystallized From 80% Ammonium Sulphate**
Exp. Method **X-ray Diffraction**

Try the Structure Explorer page for 1RNO from the new, reengineered RCSB PDB Web site!

Figure 5.5 Retrieval of protein structure from Protein Data Bank
The atomic coordinates of protein three-dimension structures can be retrieved from PDB (http://www.rcsb.org/pdb) by entering PDB id (e.g. 1RNO) or keyword (e.g. pancreatic ribonuclease). The query search returns a list of hits from which the desired PDB file can be selected (click Explore), then viewed (select View structure as shown), analyzed (select Structure neighbor, Geometry or Sequence detail) and retrieved (select Download/Display file).

TABLE 5.4 Regular secondary structures of proteins

	Optimal dihedral angles			
Secondary Structure	ϕ	ψ	Residues per turn	Translation per residue (Å)
Right-handed α helix (3.6_{13})	–57	–47	3.6	1.50
Right-handed 3_{10} helix	–49	–26	3.0	2.00
Right-handed π helix (4.4_{16})	–57	–70	4.4	1.15
2.2_7 Ribbon	–78	+59	2.2	
Left-handed α helix	+57	+47	3.6	1.50
Collagen triple helix	–51	+153		
Parallel β sheet	–119	+113	2.0	3.2
Antiparallel β sheet	–139	+135	2.0	3.4

Notes: 1. For a fully extended chain, $\phi = \psi = +180°$
2. Collagen chains can be $\phi = -51, -76, -45°$ and $\psi = +153, +127, +148°$.

residues along a chain to collect at the locations of the minima, i.e. strings of successive residues often lie in the same potential energy well. The resulting distribution means that proteins contain uninterrupted segments in which the residues have nearly the same (ϕ, Ψ) angles, and are thus either α-helices or β-strands.

5.5.1 α-Helical structure

Residues in α-helices cluster near (ϕ, Ψ) = (–60, –45) in the region I of the conformation map. Thus α helices in proteins are found when a stretch of consecutive residues are all having ϕ and Ψ pairs approximately $-52 \pm 5°$. The intermediate chain folding represented by secondary structures has two important consequences: it buries a considerable number of backbone carbonyl and amide groups, joining them in hydrogen bonding configurations, and exposes a highly textured surface composed almost entirely of side chains. Thus the innermost interactions in a folded protein find suitable intramolecular partners for a major proportion of the highly solvated peptide functional groups in a protein, without forcing ϕ or Ψ outside the favored regions for α and β secondary structure. The textured exterior surfaces ensure strong van der Waals interactions with those of complementary secondary structures elsewhere in the sequence.

Helical structures in proteins are denoted by n_x where n represents the number of residues per turn and x indicates the number of atom in the hydrogen bonded helical ring. The α-helix has the following properties:

- The hydrogen bond is formed between the N—H group of an amino acid residue n and the C═O oxygen of residue n – 4. The ring formed by a single hydrogen bond contains 13 atoms. A single turn of the helix contains 3.6 residues and therefore the α-helix is also called the 3.6_{13} helix (Figure 5.6A).
- A residue in the helix extends to1.5 Å and therefore a single turn of the helix extends 5.4 Å along the helical axis.

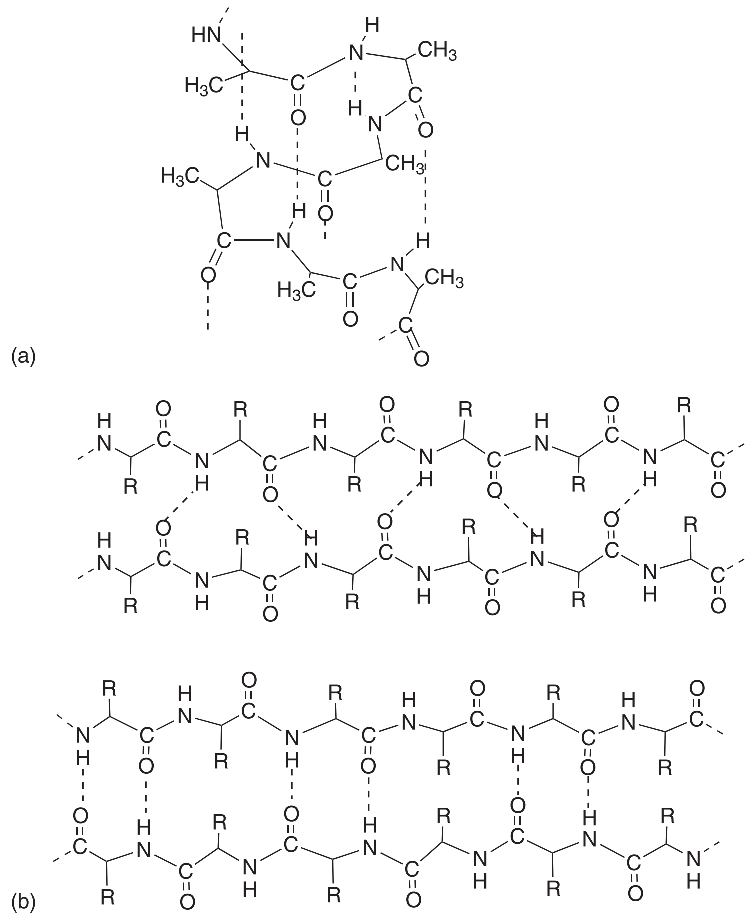
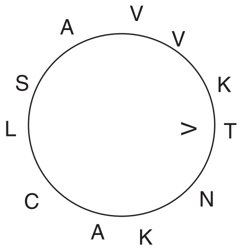


Figure 5.6 Sketches for α-helical and β-pleated structures
α-Helical (A) and β-pleated (B) structures of proteins are schematically represented. They are not drawn to dimension in order to portray hydrogen bonds in α-helix (4-residues interval and 13-atoms ring), parallel β-pleat (zigzag, equal distance) and antiparallel pleat (parallel but alternate unequal distance) formations.

- The C═O bond is parallel to the helical axis and the linear hydrogen bond formed between the C═O and N—H in the α-helix is the most stable geometrical arrangement.
- The pitch of the helix can be right-handed or left-handed, but a right-handed helix is more stable by the lesser interactions between the β-carbon and carbonyl oxygen and among the constituent atoms of the main chain.
- All the hydrogen bonds point in the same direction so the peptide units are aligned in the same orientation along the helical axis. Since a peptide unit has a dipole moment arising from the different polarity of N—H and C═O, these dipole moments are also aligned along the helical axis. The overall effect is a significant net dipole for the α-helix that gives a partial positive charge at the amino end and a partial negative charge at the carboxyl end of the α-helix.
- The peptide dipole is aligned nearly parallel to the helix axis and the strength of the electric field created by the dipole increases up to a helix length of about 10 Å (two turns). Thereafter further elongation in the helix has only a marginal effect.

In addition to the α-helix, the 3.0_{10}-helix and π-helix (4.4_{16}-helix) also exist as helical structures of the polypeptide chain. Helical structures vary considerably in length in globular proteins. ranging from 4–5 amino acid residues to over 40 residues, with an average length of ~10 residues corresponding to three turns. High content of α-helix is found in some fibrous proteins such as keratin, myosin epidermin and figrinogen, furnishing some degree of elasticity to these proteins. α Helices are also found spanning across the membrane in membrane proteins. Another type of helix occurring often in proteins is the 3_{10} helix, which has a hydrogen bond between N—H of n residue and C═O of n − 3 residue. It has a distinctive triangular appearance in the end view of the helix. In the 3_{10} helix, the α-carbons on successive turns are exactly in line with one another since there are an integral number of residues per turn, causing the hydrogen bonds to tilt relative to the helix axis. The 3_{10} helix often forms the last turn at the C terminus of an α-helix.

A convenient way to illustrate the amino acid sequences in α-helices is the helical wheel. For example, a membrane-spanning helical sequence, TASVNCAKKLV can be represented by a helical wheel as:



Since one turn in an α-helix consists of 3.6 residues, each residue can be plotted every 360°/3.6 = 100° around a circle or a spiral. Such a projection of the position of the residues on a plane perpendicular to the helical axis shows the helical wheel (spiral) with properties (e.g. charge characteristics, hydrophobicity) of α helices resulting from the distribution of amino acid residues.

5.5.2 β-Sheet structure

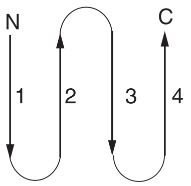
A β-Structure is built up from a combination of several regions, usually 5–10 residues of the polypeptide chain with dihedral (ϕ Ψ) angles near to −120° and 140° in the allowed upper left-hand quadrant of the conformation map (Ramachandran plot). Both interchain and intrachain β-structures exist. The interchain β-structure is found in fibrous proteins such as silk fibroin, and intrachain β-structure is present in many globular proteins. The polypeptide chains that form the β-structure are called β-strands. The β-strands are aligned adjacent to each other such that hydrogen bonds can form between N—H groups of one β-strand and C═O groups of an adjacent strand and vice versa alternatively to the left and to the right. Therefore the basic unit of β structure is not the individual β strand but the β strand pair, which form hydrogen bonds between them. The β-structure that forms from several such β-strands is pleated, termed β-pleated sheet (β-sheet), with C^{α} atoms successively sticking slightly above and below the plane of the β-sheet.

The β-strands in the β-sheet interact in two ways, i.e. parallel in which all amino acids in the aligned β-strands run in the same direction, or antiparallel in which the amino acids in successive strands run in opposite directions. In antiparallel β-sheets, a common occurrence is the Greek key topology. The adjacent β-strands in antiparallel β-sheets are often those sequential strands in the primary structure and are connected by β turns to form a β-meander. Hydrogen bond networks that hold β-strands are different in parallel and antiparallel sheets. The parallel β-sheets have evenly spaced hydrogen bonds that angle (zig-zag) across between the β-strands, whereas antiparallel β-sheets have parallel narrowly spaced hydrogen bonds that alternate with widely-spaced pairs (Figure 5.6B). Parallel β-sheets are usually buried on both sides, so that their central sequences are highly hydrophobic, and hydrophilic residues concentrate at the ends. However, antiparallel β-sheets typically bury one side in the interior and expose the other side to the solvent so that the amino acid types tend to alternate hydrophobically and hydrophilically.

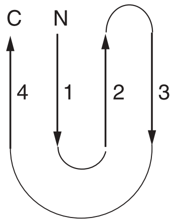
β-Strands can also combine into mixed β-sheets, with some β-strand pairs parallel and some antiparallel. The minimized energy of a parallel β sheet is higher than that of the corresponding antiparallel β sheet, indicating that a parallel β sheet is intrinsically less stable. Purely parallel sheets are least frequent, whereas purely antiparallel sheets are most common. Parallel sheets generally have at least four strands, while antiparallel sheets often contain two or three strands. Mixed sheets usually consist of 3–15 strands. Adjacent strands in an antiparallel sheet tend to be the strands that are also adjacent in the primary structure. Almost all β-sheets occurring in known protein structures have their strands twisted right-handedly at 0°–30° between strands (Sternberg and Thornton, 1978). Twists somewhat relieve the close contact of side chains directly opposite one another on neighboring β strands. An extra residue is often present in a β-strand at the edge of a sheet, interrupting the hydrogen bond network to produce a β-bulge. A β-bulge can be considered as the insertion of an extra residue into one strand, so that between a pair of hydrogen bonds there is now one residue on the normal strand but two residues on the bulged strand. Bulges are common in antiparallel β structures but rare in parallel β structures. In addition to the ubiquitous right-handed twist, β-sheets can either be curled along the strand direction or arched perpendicular to it. Curl is most common for two-stranded antiparallel ribbons. Such a β ribbon promotes compact, stable structures if its hydrophobic side chains are on the concave, close-pair side.

The most characteristic features of β-sheets are the number of strands, their relative directions (parallel versus antiparallel) and connectivity (how the strands are connected). This information can be conveyed through a simple diagram of connected arrows, known as a topology diagram for β-sheets, which may reveal some functional properties.

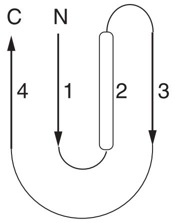
The topologies for antiparallel β-strands (from N→C) are +1, +1, +1 for β-meander, +1, +1, −3 for β_4-Greek key and +1X, −2 for βα-Greek key:



β-meander



β_4-Greek key



βα-Greek key

where arrows represent β strands and X is added for crossover connection. An α helix (cylinder) serves as the crossover connector between the first and second β strands of the βα-Greek key motif.

5.5.3 Nonrepetitive structure: Connection (loop) and turn

The right-handed helices and extended β strands are the only protein conformations in which the same ϕ, ψ angles repeat for each consecutive residue as regular, repetitive structures. The remaining portions of protein structures are made up of well-ordered but nonrepetitive conformations referred to as coils. Nonrepetitive structure involve both backbone hydrogen bonds and frequent side chain-to-main chain hydrogen bonds. The reliance on specific side chain interactions is characteristic for nonrepetitive structures. The nonrepetitive structure consists of two general types; connection or loop and turn.

Most protein structures are built up from combinations of α helices and β strands, which form stable hydrophobic core, and connected by loop regions. Connections or loops range from large, compact Ω loops (Leszczynski and Rose, 1986), which have close termini, good internal packing interactions and many side chain-to-main chain hydrogen bonds, to relatively extended straps for getting from one piece of repetitive structure to another. At the surface of the molecules, globular proteins often reverse the direction of polypeptide chains via the tight turns (reverse turns). Turns are the shortest possible connections/loops involving in the direction changes or breaks between other pieces of structures. Various types of reverse turns occur, which reverse the direction of the polypeptide chain by 180°. The reverse turns consisting of three residues (i.e. residues, i to i + 2, with one residue not involved in the hydrogen-bonding network) are termed γ turns, and those with four residues (i.e. residues, i to i + 3, with two residues not involved in the hydrogen-bonding network) are known as β turns. The β and γ turns are compared in Table 5.5. β Hairpins refer specifically to the reverse turns linking adjacent strands in an antiparallel β-sheet.

The conformations of turn regions depend primarily on the positions of certain amino acid residues (usually Gly, Pro or Asn) in the loop (Hutchinson and Thornton, 1994). The type I β turn can accommodate any amino acid at positions i through i + 3, except that Pro cannot be at position i + 2. For the type I and II β turns, Gly and Pro predominate at positions i + 3 and i + 1 respectively, whereas Asn, Asp, Cys and Ser occur frequently at position i. The type II β turn shows preference for Gly and Asn at position i + 2. Gly occurs at positions i + 1 and i + 2 in type I′ as well as type III′ turns, and i + 1 in type II′ turns. The conformation of the type III turn corresponds to that of one turn of the 3.0_{10}-helix, while Type VIa and VIb have a *cis* peptide bond.

TABLE 5.5 Comparison of γ and β turns

Bend type		Dihedral angles of central residue(s) (deg)			
		ϕ_{i+1}	ψ_{i+1}	ϕ_{i+2}	ψ_{i+2}
γ:	classical	70 to 85	–60 to –70		
	inverse	–70 to –85	60 to 70		
β:	I	–60	–30	–90	0
	I’	60	30	90	0
	II	–60	120	80	0
	II’	60	–120	–80	0
	III	–60	–30	–60	–30
	III’	60	30	60	30
	IV	Any bend with two or more angles differing by >40° from those given			
	V	–80	80	80	–80
	V’	80	–80	–80	80
	VIa	–60	120	–90	0
	VIb	–120	120	–60	0
	VII	Kink in chain			
	VIII	–60	–30	–120	120

5.5.4 Notes to secondary structures of globular proteins

The following tendencies have been observed for secondary structures and their connections in globular proteins:

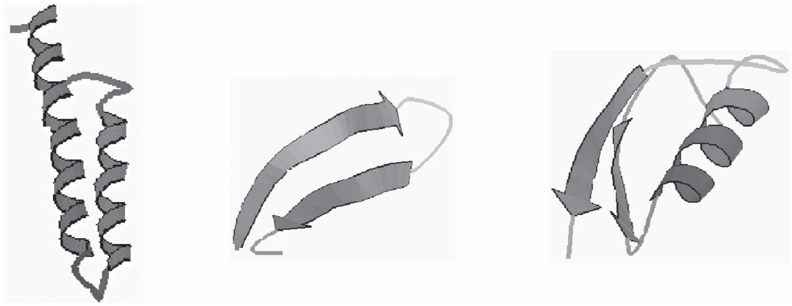
- Secondary structure is most apparent in large globular proteins where it comprises most of the interior to provide the required stability of the folded conformation.
- A single α-helix usually consists of 10–15 residues. Almost all of the helical conformations found in proteins are α-helices. 3.0_{10}-helices seldom present in protein structures. Observed 3_{10} helices in proteins are rarely longer than one turn (Karpen *et al.*, 1992), and they occur most often at the ends of α-helices where they serve to change direction. They are sometimes found at the carboxyl end of an α-helix, but do not exist at the amino end of an α-helix. In the 3.0_{10}-helix, a straight hydrogen bond is not formed between N—H and C═O groups, and thus the helix is not as stable as the α-helix. Although the straight hydrogen bond between N—H and C═O groups is formed in π-helix, there is a large space around its helix axis and van der Waals interactions are not utilized to stabilize the helix.
- The left-handed α-helical conformation provides a potentially important means of changing the direction of a polypeptide chain (in addition to reverse turns). This conformation has a slightly less favorable energy than that for right-handed helices. However, the rarely observed occupants of this structure (about 1% of all residues) are invariable single residues. These α_L residues occur at the carboxyl termini either of α-helices or β-strands, abruptly terminating these secondary structures and introducing substantial changes in direction, e.g. in cytidine deaminase (Betts *et al.*, 1994). The other frequent location of α_L conformations is in position 1 of the β-bulge (Chan *et al.*, 1993; Richardson *et al.*, 1978).
- Two α-helices form an α-hairpin and two β-strands form an βXβ turn, where X is an α-helix or coil. The connection of the two β-strands in virtually all βxβ turns is right-handed.
- Helices have the general shape of cylinders. They pack together with one face on the interior of the protein and the other exposed to the solvent.

- In solvent exposed α-helices, the plane of the peptide bond is often rotated so that the C═O group points outward from the helical axis toward the solvent. The helical axis is often curved with the surface on the outside of the globular structure somewhat extended.
- Single β-strand is usually 3–10 residues long. Most β-pleated sheets consisting of several β-strands are not planar but twisted, known as a β-twist corresponding to a right-handed rotation of carbonyl and amide groups. Various mechanisms augment the curvature in naturally occurring β-structures. The most important of these entails the inclusion of two residues between two β-type hydrogen bonds, known as a ‘β-bulge’ (Richardson *et al.*, 1978), which is nearly always found in highly twisted β-structures.
- β-Pleated sheet structures are actually a special case of secondary structure because they cannot form without intervening secondary structure, which usually take the form of an α helix. The most common such structure is ‘βαβ crossover connection’ (Richardson, 1973), where the two connected β-strands are adjacent.
- Most parallel β-sheets are protected by an α-helix or other structure. In the antiparallel β-sheet, one side of the sheet is exposed to solvent and the other is buried in the protein interior. Thus a hydrophilic residue and hydrophobic residue appear alternately in the amino acid sequence. The parallel β-sheet seems to be less stable than the antiparallel β-sheet due to the non-linear hydrogen bonds.
- Types I, II and III reverse turns are closely related. From their dihedral angles it can be seen that the Type I turn begins roughly in the region of the right-hand α-helix and ends close to the region occupies by a β-strand, whereas Type III is actually a helical configuration (3.0_{10}).
- The reverse turns are usually located on the surface of the protein molecule. In addition to their function as connecting units between secondary structural elements, loop regions frequently participate in forming binding and catalytic sites. Turns can become a dominant element in small proteins, especially those that bind metals. In small metalloproteins, turns serve to create internal volumes for the metal ions or clusters and to provide NH—S hydrogen bonds to cysteine S_γ atoms that bind the metals.
- The regions of polypeptide chain, which do not form regular secondary structures, are referred to as coil conformation (random coil).
- Pieces of secondary structure that are adjacent in the sequence are often in contact in three dimensions and usually pack in an antiparallel manner.
- The connections between secondary structures neither cross each other nor make knots in the chain.

5.5.5 Motifs: Supersecondary structures

Certain groupings of secondary structure elements, including the segments of polypeptide chain that connect these structural elements, occur in many globular proteins and are termed motifs or supersecondary structures (Figure 5.7). The motifs are a higher level of organization that preserves the structure hierarchy. These are:

(1) Helix-turn-helix (αα) motif, also known as αα hairpin in which two successive antiparallel α helices pack against each other with their axes inclined so as to permit intermeshing of their contacting side chains, giving rise to the stable coiled coil conformation. This motif is specific for DNA binding or Ca^{2+} binding.



αα motif ββ motif βαβ motif

Figure 5.7 Common motifs of protein structures

(2) Hairpin β motif, also called ββ hairpin, which consists of two sequential antiparallel β-strands connected by a tight reverse turn. This motif changes the direction of antiparallel β-sheet structures.

(3) β Barrel motif in which extended antiparallel β sheets roll up to form a barrel with geometric pattern similar to the ornamental Greek key (topology: +3, −1, −1) on ancient Greek pottery or β-mender (topology: +1, +1, +1, +1) from the native American basket.

(4) βαβ Motif in which the right-handed crossover of two consecutive parallel β-strands are connected by an α helix. Just as the strands of the β-structure are always twisted in a right-handed sense, so the common βαβ crossover nearly always makes a right-handed spiral rotation. The common structure is a pair of βαβ crossover connections oriented approximately 180° apart, such that the four β-strands form a single parallel β-sheet. This motif forms nucleotide binding site domains of dehydrogenases and kinases. It is the fundamental active site building block for a statistically significant fraction of all enzymes. Representative enzymes include examples from intermediary metabolism (Faber and Petsko, 1990), amino acid biosynthesis (Wilmanns *et al.*, 1991), translation (Carter, 1993), energy transduction (Abrahams *et al.*, 1994) and signal transduction (Pai *et al.*, 1990).

Sometimes motifs are also used interchangeably with patterns to describe recurring, conserved sequences of functional significance derived from sequence/structure alignment studies. In this usage, motifs (patterns) are commonly associated with specific functions of homologous proteins (Table 5.6).

5.6 DOMAINS AND TERTIARY STRUCTURES OF PROTEINS

Complementary packing surfaces are the key link to the next hierarchical level, tertiary structure. Tertiary interactions arise because distant parts of a protein interact across what might be called an intramolecular interface. There is a consensus that tertiary structure entails at least two new concepts:

1. The main chain adopts a characteristic fold with some degree of imprecision, both with respect to primary structure encoding (Jones *et al.*, 1992; Srinivasan and Rose, 1995) and to the mechanism of renaturation (Dill and Chan, 1997).
2. Packing or stereochemical complementarity (Richards and Lim, 1993) is an important source of the new bonding interactions, required to stabilize unique native structure.

TABLE 5.6 Examples of sequence motifs/patterns

Motif	Function	Example	Remark
CxxCH	Heme attachment	Cyt c3, c551, c′; Cyt f	This motif binds covalently to a heme group via two Cys and coordinates the bound iron with His of many homologous cytochromes.
CxxxxxRS	Phosphatase (type I/II)	Type I and II Pases	The motif is contained within a loop connecting a β-strand to an α-helix around the active site.
GxxGxxKT	P-loop	Ras-p21-GTPase, PEP carboxykinase, uridylate kinase	P-loops occur in many doubly wound α/β structures. This motif functions by binding the phosphate backbone of a mononucleotide.
GxGxxG	FAD/NAD binding	PFK, LactateDH, *iso*citrate DH, glucose oxidase	FAD/NAD binding motif within doubly wound α/β protein, adopts a conformation suited to bind the phosphate backbone of dinucleotides.
CxxCG	Zn finger	Protein kinase C, AdK, HIV NCP, raR	This motif is found in zinc-containing small, cysteine-rich domains and it is central to their core structure.
CxxCxxC	Fe-S binding	Ferredoxin, TMA DH	Three Cys are involved in two turns of an α-helix, in addition to a four-residue loop that surrounds a single Fe-S cluster.
SxDGxxW	Asp box	Chitobiosidase, neuraminidase, bacterial RNase	This motif is found in at least three different folds; immunoglobulin fold, β-propeller and antiparallel β-sheet with a single α-helix.

Notes: 1. Taken from Lupas *et al*. (2001).

2. Abbreviations used: AdK, adenylate kinase; Cyt, cytochrome(s); DH, dehydrogenase; Pase, phosphatase; NCP, nucleocapsid protein; PEP, phosphoenolpyruvate; PFK, phosophofructokinase; raR, DNA binding domain of retinoic acid receptor; RNase, ribonuclease(s); TMA, trimethylamine; x, any amino acids.

Qualitatively, tertiary interactions include the same type of interactions that stabilize secondary structure: intramolecular hydrogen bonds, intramolecular hydrophobic bonds and to a variable extent, electrostatic interactions between charged groups. Formation of tertiary structure converts a subset of these surfaces into inside surfaces. Proteins do achieve an extraordinary degree of complementarity between distant parts. In general, tertiary interactions tend to involve a higher proportion of nonpolar side chain packing, relative to the number of intramolecular hydrogen bonds between main chains. Imperfections in stereochemical complementarity can and are compensated by including buried water molecules wherever there are a sufficient number of accessible hydrogen-bonding sites (Radzika and Wolfenden, 1994; Zhang and Hermans, 1996). Moreover side chain-side chain and side chain-main chain hydrogen bonds may contribute.

Quantitatively, tertiary interactions fold polypeptide chains into one or several domains to form compact globular structure. Their compactness, which characterizes domains, can be expressed as the ratio of their surface area to the surface area of a sphere versus the same volume with an observed values of 1.64 ± 0.08. Thus the domain concerns a polypeptide chain or a part of a polypeptide chain that can independently fold into compact, stable tertiary structure.

5.6.1 Domain structures

The folded structures of most small proteins are compact and reasonably spherical. However, the structures of most proteins with more than 200 amino acid residues appear to consist of two or more structural domains. A domain can be defined as that part of a protein that can fold up independently of neighboring sequences. Structurally, a domain is viewed as a compact, spatially distinct unit. Biochemically, domains are described as protein regions with assigned experimental functions. In sequence comparison, domains are viewed from an evolutionary perspective and described as sequence-similar homologs that are often present in different molecular contexts. These views are compatible for domains that refer to sequence-similar homologs adopting similar folds and possessing comparable functions (Ponting and Russell, 2002). Thus domains that are formed by different combinations of secondary structure elements and motifs are also units of functions. Many active sites occur at domain interface. In enzymes with more than one substrates/effectors, the different binding sites usually occur on different domains. In proteins with more than one function, different domains usually perform the different functions. The basis of domain organization is the cohesion between side chains that stabilizes unique structures. As such, there is a need for a critical number of residues in a sequence. A single domain usually consists of 100–150 amino acid residues and is about 25 Å in diameter. The domains in a given protein may be packed together tightly or loosely. If they are loosely packed, domains can be dissected by limited proteolysis, providing suitable target peptide bonds are available in the connecting segments. Each proteolytic fragment has the same conformation as that existing in the native protein molecule.

The domains (interchangeable used as structure classes here) can be classified as:

5.6.1.1 α Domain structures. α Domain structures consist of a bundle of α helices that are connected by loops on the surface of the domains. All helical proteins are generally either antiparallel or nearly perpendicular, due presumably to a preference for near-neighbor helix dipoles or helix-to-helix packing in which the ridges and grooves on the helix surface intercalate. The α helices are packed against each other so that one side of each α helix provides the hydrophobic side chains for packing interactions in the interior core and the other side faces the solvent. Two most frequently encountered structures are the four-helix bundle in which each pair of sequentially adjacent α helices are joined in an antiparallel fashion and the globin fold in which a bundle of eight α helices wrap around the core in different directions. The four-helix bundle is the simplest and common type of antiparallel α structure in an approximately cylindrical cluster of helices with an up-and-down, near-neighbor connectivity. The globin fold structure is based on successive helix pairs that are close to perpendicular and touch at their sequence-adjacent ends. The great majority of adjacent helix pairs are arranged with the two helix axes intersecting near their common axes as though they had formed by bending a single, longer helix.

5.6.1.2 β Domain structures. The cores of these domains are built up by 4–10 antiparallel β strands arranged in a predominantly antiparallel fashion to form two β sheets that are joined together and packed against each other. The β sheets are twisted to form a barrel-like structure. The three most often encountered groups are:

1. *Up-and-down β barrels* in which successive β strands are added adjacent to the previous strand until the β barrel is closed (eight antiparallel β strands in β barrels or six small sheets, each with four β strands in superbarrels).
2. *Greek key barrels* in which eight antiparallel β strands builds up Greek key motifs (two consecutive Greek key motifs with one of the connections crossing one end of the barrel).
3. *Jelly roll barrels* in which any even number of β strands (greater than four β strands) form a jelly roll barrel with equal number of connections across the top and bottom of the barrel. The antiparallel up-and-down β barrel is the simplest organization,

having all +1 near-neighbor hairpin connections between the strands. The most common β structure is the Greek key barrel with +3 hairpin connection. One way of characterizing the Greek key β-barrel is the presence of a long pair of superstrands that follow next to each other. The swirl of this superstrand pair is always counterclockwise when viewed from solvent. The chief features that vary from one Greek key topology to another are the number and arrangement of extra up-and-down strands added on either side of the Greek key core. All jelly roll β-barrels provide the covering helices or loops either within the β sheet region of the sequence or immediately adjacent to it. The covering is sometimes made up of right-handed crossover connections and sometimes from the long extension of a hairpin connection.

5.6.1.3 α-β Domain structures. Domains with α helices and β strands are divided into α/β domains, which have mixed or approximately alternating segments of α helical and β strand secondary structures, and α + β domains in which the α helical and β sheet regions are somewhat segregated. The most common and regular domain structures are built from βαβ motif to form α/β domains, which consist of central β sheets surrounded by α helices. In these structures, the core β sheets and α helices, which stabilize the tertiary structure, are connected by loop regions. The dominant organizing principle of parallel α/β proteins is the overwhelming right-handedness. All the glycolytic enzymes are α/β domain structures. Two main types of α/β domain structures are parallel β barrel and open twisted β. In parallel β barrel, a core of eight twisted parallel β-strands are arranged into a barrel with the connecting α helices on the outside of the barrel. In open twisted β, the βαβ motifs are connected to form an open twisted β sheet surrounded by α helices on both sides of the β sheet.

In all parallel β barrel structures of the α/β domain enzymes, the active site is situated in the bottom of a funnel shaped pocket corresponding to the eight loops that connect the carboxyl end of the β-strands with the amino end of the α helices. Residues that participate in binding and catalytic activities are in these loop regions. In some structures, an additional loop region from the second domain or a different subunit comes close to the active site to participate in binding and catalysis. The open twisted β domains are found in the nucleotide-binding domain of dehydrogenases and kinases. The open β sheet generally consists of 4 to 10 β-strands and it can be a mixed β sheet where hairpin connections give rise to some antiparallel β-strands mixed with the parallel β-strands.

5.6.1.4 Miscellaneous. Irregular domains consist of small number/mixed but isolated α helices and β strands or small metalloproteins and sulfide containing proteins, which do not fall into either one of the α, α/β or antiparallel β domains. These structures have relatively small amounts of secondary structure; what they do have is often irregular and not organized but neatly arranged into larger arrays. The disulfide bonds or bound metals may substitute for the role of extensive secondary structure in stabilizing the folded proteins.

5.6.2 Tertiary structures and protein folds

Protein folds have been used interchangeable with protein topologies. Topologically, protein chains show clear regularities such as:

- Secondary structures in proteins usually pack in one of a small number of relative orientations.
- Segments of secondary structures that are adjacent in the sequence are often in contact in three dimensions and usually pack in antiparallel. Three common structures are αα (two antiparallel packed α helices), ββ (two antiparallel-packed β-

strands), βαβ (one α helix packed against two adjacent β-stands), whilst αββ and ββα folding units are rare occurrences.

- The connection in β-X-β, where β's are parallel strands in the same sheet and X is an α helix or a loop in the different sheet and is right-handed.
- The connections between secondary structures neither cross each other nor make knots in the chain.

The criterion for attributing a particular family name to a new structure is a property of protein tertiary structure that invites systematization and has come to known as the fold. The inflexibility of the polypeptide chain and side chain interactions imposes restriction on the limitation of favorable folding patterns for secondary structures to form tertiary structures.

Proteins with related folds fall into one of four broad grouping, depending on the constituent secondary structures and how they are combined (Levitt and Chothia, 1976). These are designated α, β, α/β and α + β, to indicate proteins with all α, all β, interleaved α and β, and segregated α and β structures. The designations also apply to parts of larger proteins built from domains that fit one of the four preceding categories. Proteins in the first two categories are relatively easy to identify because they consist of a single secondary structural element. The remaining categories are less clearly defined, so there is a need to provide criteria on which to classify them. There is some variation in how these criteria are formulated. A scheme proposed by Chou (Chou, 1995) is described in Table 5.7. Three major categories of folds are described below:

α Helical folds. In helix folding (packing) of two α helices, the distance between the helix axes varies between 6.8 Å and 12.0 Å, with the mean interaxial distance of 9.4 Å (e.g. myoglobin, cytochrome C). The relative orientation of the two α helices packed face-to-face can be expressed by the angle (Ω) between the helix axes. The values for this angle in observed helix packing cover the whole possible range with a major peak at $\Omega = -50°$ and a minor peak at $\Omega = +20°$. Residues on the surface of helices form ridges separated by grooves. Helices fold together by the ridges of one helix packing into grooves of the other. The ridges on the helix surface are formed by residues whose separation is usually four (±4n ridges) and occasionally three (±3n ridges). Folding helices incline the axial angles of ~50° when both helices use ±4n ridges, and ~20° when one helix uses ±4n ridge and the other ±3n ridge.

Four helix bundles are a common structural motif that can be observed independently as well as components of large folding units. The six classes of four helix bundles constructed from four overlapping sets of helix pairs (where the helical axials are ~+20° (||) and ~−70° (⊥) respectively) are:

1. square (||, ||, ||, ||) e.g. apolipoprotein E3,
2. splinter (||, ||, ||, ⊥) e.g. A-chain of granulocyte-macrophage colony stimulating factors,

TABLE 5.7 Classification of protein tertiary structures

Class	Characteristics
α proteins	$\alpha \geq 40\%$, $5\% \geq \beta$
β proteins	$\beta \geq 40\%$, $5\% \geq \alpha$
α/β proteins	Interleaved $\alpha \geq 15\%$, $\beta \geq 15\%$ with > 60% of strands parallel β
α+β proteins	Segregated $\alpha \geq 15\%$, $\beta \geq 15\%$; with > 60% of strands antiparallel β
ξ proteins	$10\% \geq \alpha$, $10\% \geq \beta$

3. x (||, ⊥, ||, ⊥) e.g. 3-isopropylmalate dehydrogenase,
4. unicornate (||, ||, ⊥, ⊥) e.g. T4 lysozyme,
5. bicornate (||, ⊥, ⊥, ⊥) e.g. neutral protease, and
6. splayed (⊥, ⊥, ⊥, ⊥) e.g. A-chain of sarcoplasmic calcium binding protein

The patterns for ideal helix assemblies are:

a) Helices have the general shape of cylinder.
b) Helices pack together with one face on the interior of the protein and the other exposed to the solvent.
c) Helices pack around a central hydrophobic core whose diameter is about the length of two residues (~11 Å).
d) Helices are close packed with a similar number of contacts.
e) Assemblies of helices must be as spherical as possible.

β Sheet packings. Parallel β sheets usually have both faces covered by α helices. Antiparallel β sheets are usually folded against one another. β Strands are packed in different ways, which allow β sheets to twist, coil and bend (e.g. chymotrypsin, immunoglobin). Three distinct classes of β sheet folds are observed:

1. β Barrel in which parallel β sheets form the core with one or both faces covered by α helices in α/β structures or the antiparallel β sheet twists and closes itself in a cylinder.
2. Aligned β packing in which β-strands in different sheets are linked by loops and right-handedly twist, resulting in the main chain angle between the two packed β sheets with an angle of ~−30°.
3. Orthogonal β packing in which β sheets folds upon themselves to form layer structures with an inclination of ~90° between the two layers.

The β-strand at two diagonally opposite corners passes from one layer to the next without interruption. The bend in the β packing is usually produced by local coiling of the strand, a β-turn of a β bulge. The large majority of the known assemblies of β sheets are two-layer structures with limited chain topologies, namely hairpin, Greek key and jelly roll for two, four and six strands respectively.

Mixed packings of α helices and β strands. Protein structures (e.g. alcohol dehydrogenase, triosephosphate isomerase) composed of α helices and β strands that roughly alternate along the peptide chain (α/β), can be folded into two classes:

1. The β sheet is just twisted and the α helices pack upon its face.
2. The β sheet is twisted and coiled upon itself to form a barrel structure with helices packed on its outside surface.

Ideal α helices fold on to ideal β sheets with their axes parallel to the strands. Some small α/β proteins contain just two layers, i.e. α helices packed on one side of a β sheet. However, the large majority of proteins have three layer structures with α helices packed on both sides of a β sheet. In the α/β barrel, a β sheet coils round to form a closed cylinder. The barrel structures are formed by eight parallel strands of β sheets, which are linked by α helices with their axes approximately parallel to the strand direction.

The motif βαβ, which is prominent in the α/β proteins, is absent in protein α + β structures (e.g. lysozyme, papain) with separate α helices and β strands in different parts of structures/domains. An appreciative amount of α helices and/or β sheets can be treated as their respective folding patterns.

Proteins built up entirely from α-helices include the globin and hemerythrin families and bacteriorhodopsin. These are transport proteins and membrane-bound proteins. All-β proteins invariably consist of a sandwich of two layers, each a sheet held together by a hydrophobic core formed by one side from each sheet. Often the sheet is rounded to form a β-barrel. The β interactions in these proteins are usually antiparallel because there are no helices available to form the necessary crossover connections required for parallel β-sheets. The novel family of proteins with an all-parallel β-helix (Jurnak *et al.*, 1994; Yoder *et al.*, 1993) is a remarkable exception. Proteins with distinct regions of all-α or all-β structures are called α + β proteins. The most complex class is the α/β proteins. In these, segments of α-helix and β-sheet follow one another, and consequently each of the two classes of secondary structure is intimately involved in stabilization of the other. For this reason, the β-structures are often parallel, since the presence of helices provides the necessary crossover connections. The ζ-class of proteins is an ill-defined catchall (Chou, 1995). It includes a number of large peptides with no globular structure.

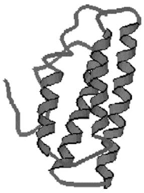
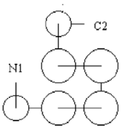
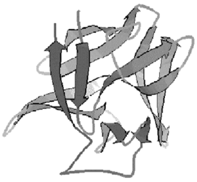
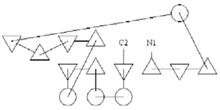
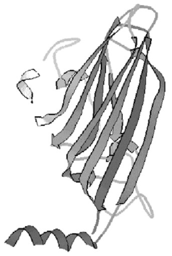
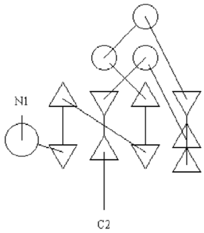
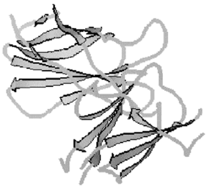
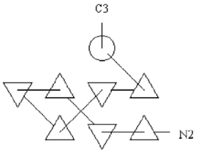
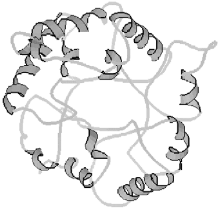
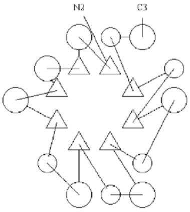
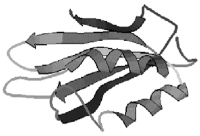
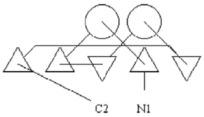
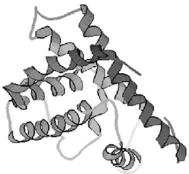
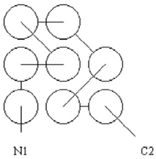
The progression of strands along a β-sheet changes direction when a connecting loop skips one or more strands. This motif was dubbed the Greek key (Richardson, 1977). Another important example of changes in direction arises in proteins with a single parallel β-sheet. It is important to protect both sides of the sheet from dissolving in water, and to accomplish this the direction must change at least once. Otherwise all of the crossover connections would be made on only one side of the sheet. This is probably the reason when the α/β proteins have so many structural similarities and so few distinct families. Either the helices all lie on the same side, in which case the β-strands are protected by forming the observed $(\beta\alpha)_8$ TIM (triosephosphate isomerase) barrel side-wise (Nobel *et al.*, 1991), or else the helices lie on opposite sides and the protein has a single sheet.

An unexpectedly large number of the proteins in each class exhibit strong internal symmetry, whereby repeated structural elements are arranged in nearly symmetrical fashion, usually related by two-fold rotations such as observed in the subunit of dimeric *E. coli* cytidine deaminase (Betts *et al.*, 1994). Despite the absence of obvious sequence homology between the two halves of the enzyme, they clearly share similar folding instructions.

In extracellular proteins, the oxidizing environment leads to formation of disulfide bonds whenever two free cysteine side chains can assume the appropriate configuration. Not surprisingly, the range of tertiary structures in proteins with disulfide bonds is more varied than that normally observed for intracellular proteins.

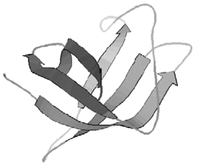
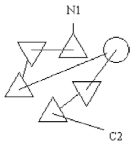
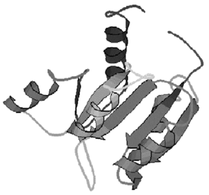
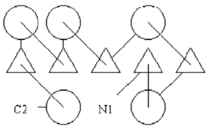
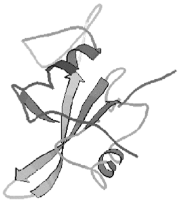
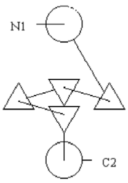
Superfolds. Despite their diversity, proteins adopt a limited number of structural folds. Proteins are classified within the same fold if they have the same major secondary structural elements in the same mutual orientation and with the same connectivity. These folds are largely formed by limited recurring supersecondary structure elements. Proteins can adopt the same topological fold despite having very dissimilar sequences. Certainly stability and folding efficiency have contributed. Some folds may also have yield better scaffolds for the establishment of active sites, displacing less-suitable folds. Most folds seem to consist of a limited number of supersecondary structures, namely ββ-hairpins, αα-hairpins and βαβ elements plus β_4-Greek key (Richardson, 1977) and βα-Greek key (Efimov, 1995) motifs. As a consequence, the distribution of structures amongst the different fold families is highly biased toward a small number of folds termed superfolds, which account for approximately one-third of all proteins of known structures (Orengo *et al.*, 1994; Salem *et al.*, 1999) as listed in Table 5.8.

TABLE 5.8 Some characteristics of superfolds

Fold	Representative structure	TOPS representation	Class	% sss	Examples
UpDown			α	90	2HMZ(A), 1RPR, 256B
Trefoil			β	83	4FGF, 1ILB, 1AAI
Jelly roll			β	47	2STV, 1GOH, 1TNF
Ig fold			β	67	3HHR(B), 1JP5, 2SOD
TIM barrel			α/β	82	7TIM(B), 1HTI, 4ENL, 1XIM, 2TAA
Plaitfold			$\alpha + \beta$	38	1APS, 1FXD, 1BOP
Globin			α	88	1EBC, 1HLM

(continued)

TABLE 5.8 ***continued***

Fold	Representative structure	TOPS representation	Class	% sss	Examples
OB fold			β	77	1MJC, 1QVC
Doubly wound			α/β	68	5CHY, 3ECA, 3PGM, 3ADK, 4DFR
UB roll			α + β	55	1LKK, 1SHA, 1UBQ

Notes: 1. Taken in part from Salem *et al*. (1999).

2. Examples are given in PDB codes. Representative structures and TOPS are from the underlined PDB files.

3. Abbreviations used: Ig, immunoglobulin-like; sss, supersecondary structures; OB, oligonucleotide/oligosaccharide binding; UB, ubiquitin-like αβ-roll.

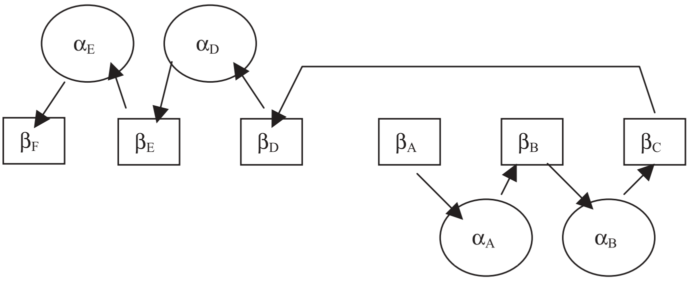
The up-down α-fold and globin fold structures are dominated by αα hairpin interactions. The trefoil fold is formed from three two-stranded β hairpins that bend in the middle. The beginning and end strands of each hairpin form a β-barrel, while the distal strands form a trigonal array capping the barrel. The triosephosphate isomerase (TIM) barrel folds consist ideally of eight repeating core βαβ motifs with peripheral surface α helices and/or β strands. The α/β motifs wind via a long loop to form the doubly-wound fold (Buehner *et al.*, 1973) to which the Rossmann fold (Rossmann *et al.*, 1974) belongs. Energy considerations have shown that the superfolds are able to support a much broader repertoire of sequences than other folds (Shakhnovich *et al.*, 2003). The regularity and optimal packing of the fold will ensure the population of these arrangements during evolution of protein structures (Orengo and Thornton, 2005).

5.6.3 Folds and protein binding

Although no firm correlation can be established between the type of fold and the function of protein, there appears to be some correlation between the ligands bound and fold type. For example, NAD(P)$^+$/NAD(P)H cofactors show a preference for binding to α/β Rossmann-like folds and DNA to mainly-α receptor-like folds. Furthermore, some functions are more frequently observed in particular structural class. For example, enzymes are much more commonly α-β proteins, while extracellular proteins are frequently mainly-β.

Conversely, DNA binding and carbohydrate binding proteins are mainly-α (Martin *et al.*, 1998).

The quintessential functional behaviors of proteins are related to their abilities to bind and change shape, and to the coupling between binding and conformational change. There is an strange association between the approximately symmetric arrangements of long stretches of polypeptide within a protein and the location of binding sites. In such cases, two polypeptide segments enter or leave a region of space in opposite directions. This phenomenon was first recognized for β-structures since discontinuities in β-structure invariably lead to places where the surface of a protein must open into a cavity. These regions have been called 'topological switch points' (Brändén, 1980), and they are a reliable indicator of ligand binding sites. The topological switch point evident in the Rossmann fold appears to be the quintessence of the ability of proteins to manage the free energy of hydrolysis of nucleotide triphosphates.



The Rossmann nucleotide-binding fold is almost always found in proteins that bind to purine nucleotide coenzymes. The long loop between β_C and β_D tends to create a natural pocket for the nucleotide ligand. The N—H groups in the last turn of the α_A helix are well positioned to form hydrogen bonds to phosphate oxygen of the ligand. In dehydrogenases, the adenine and nicotinamide moieties of $NAD^+/NADP^+$ bind to pockets more or less on opposite sides of the β sheet from the crossover α helices. In such cases, the loop between a central β-strand and α-helix invariably forms a binding site for one of the phosphates connecting the two-nucleotide moieties. A similar arrangement occurs in unrelated proteins containing the Rossmann fold, such as the class I aminoacyl-tRNA synthetases (Carter, 1993). In tryptophanyl-tRNA synthetase, the Rossmann fold binds the activated amino acid in a similar fashion, the amino acid occupying the location occupied by the nicotinamide in the dehydrogenases. Ten out of the twenty aminoacyl-tRNA synthetases use it to bind ATP for amino acid activation (Carter, 1993).

A close topological relative of the Rossmann fold, but with a somewhat different nucleotide binding mode, is highly conserved motif in kinases, G-proteins and motors (Smith and Rayment, 1996). This variant does not involve successive β-α-β crossovers, but retains only the two central β-strands, each of which is followed by an α-helix that veers away in the opposite direction. Related to the Rossmann dinucleotide fold in its structural organization, this motif invariably utilizes a conserved loop between the central β-stand and the following helix. This loop, called the 'P-loop' (Walker *et al.*, 1992), is a signature of nearly all enzymes that utilize the γ-phosphate of ATP or GTP. The consensus sequence of this loop, Gly-X-X-X-X-Gly-Lys-Thr-Ser, is adapted to bind the α-phosphate of the nucleotide close to the amino terminus of the following helix. Across the switch point in these proteins, there is invariably a region (related by the approximate symmetry to the P-loop) called 'switch II', which is crucial to the coupling between nucleotide-triphosphate hydrolysis and communication of this event to and from other protein partners.

5.6.4 Membrane proteins

Proteins, which compose 25–75% of the mass of most natural membranes, mediate various functions of membranes such as transport, catalysis, signal transduction and structural integrity. Membrane proteins have amphipathic structures that reflect the membrane in which they reside. They have both polar surfaces that interact with aqueous environment and nonpolar surfaces that interact with the hydrophobic interior of lipid bilayer. Nonintegral membrane proteins are essentially water-soluble but are anchored to the membrane only by fatty acid chains attached covalently by their polar ends to the proteins. Integral membrane proteins (referred to as simply membrane proteins) are immersed in lipid bilayer because their polypeptide chains generally traverse the membrane completely. This may involve single transmembrane chain or multiple transmembrane segments, which are connected by loops of varying size. The polypeptide chains of some membrane proteins are almost entirely within the membrane, with only a few residues exposed to the aqueous solvent. In other membrane proteins, the segment accessible to the solvent may be extensive and may correspond to one or more domains similar to those of water-soluble globular proteins.

Membrane proteins differ from water-soluble proteins in their nonpolar interface with membrane. The surface of membrane proteins in contact with the lipid bilayer is extremely nonpolar such that this surface is more hydrophobic than the protein interior. Most integral membrane proteins traverse the membrane as α-helical segments (e.g. rhodopsin, Figure 5.8A). Thus the primary structures of most membrane proteins have one or more hydrophobic segments that can form hydrophobic helices of the required length (30–40 Å corresponding to 20–27 residues/helix). However, a 16-stranded β-barrel of bacterial outer membrane protein, porin (Figure 5.8B) is an exception to the hydrophobic transmembrane helices. The barrel of porin is perpendicular to the plane of the membrane with an interior pore through the membrane, probably functioning as a selective channel for polar solutes.

5.6.5 Fibrous proteins

Most fibrous proteins have regular extended structures representing a structural complexity intermediate between pure secondary and tertiary structures of globular proteins. This conformational regularity is derived from regularities in their amino acid sequences. Table 5.9 lists some of their structural elements.

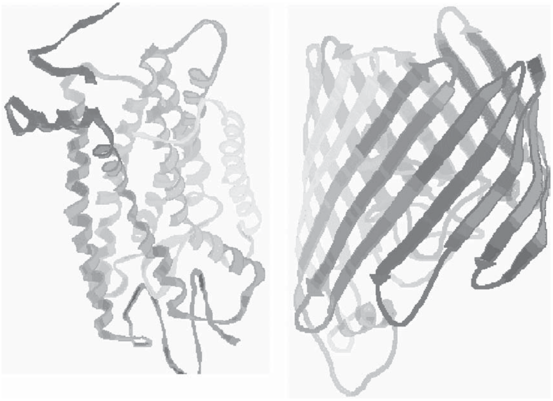
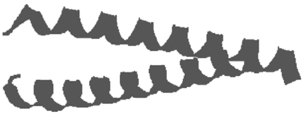
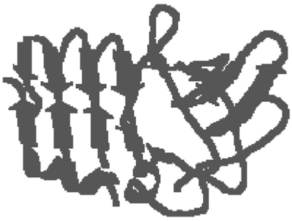
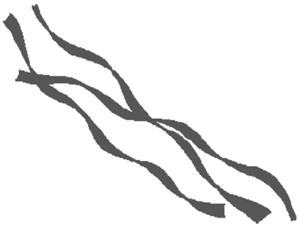

Integral membrane rhodopsin (1L9H.pdb)
Bacterial outer membrane porin (2POR.pdb)

Figure 5.8 Representative membrane folds
Two representative membrane folds are illustrated for (A) all α-helices of rhodopsin (1L9H.pdb) with seven membrane traversing α-helical segments and (B) 16-stranded β-barrel structure of porin (2POR.pdb).

TABLE 5.9 Structural elements of fibrous proteins

Structural elements	Representative fold	Characteristics
Coiled coil (1ZIK.pdb)		Amphipathic α-helices associate via their hydrophobic faces in a knobs-into-holes side-chain arrangement. α-Helices are characterized by heptad repeats (repeated seven residues, a–g, within helix). The helices interact via hydrophobic residues between residues a and d to form an apolar stripe along one side and an electrostatic interaction between residues e and g on the other side of each helix.
β-Roll (1SAT.pdb)		Right-handed β-rolls are formed by stacked coils of two or more β-strands built of short repeats. The silkworm fibroin consists of antiparallel β-sheets with ~50 repeated – $(GA)_2GSGAAG-(SGAG)_8Y$-sequences. The β-sheets are stacked on top of each other with Gly tending in contact with each other alternating with contacting Ala and Ser to give close-packed sheets.
Triplet helix (1BKV.pdb)		Three polypeptide chains of collagen are coiled together, each with a slightly twisted, left-handed, three-fold helical formation. Collagen polypeptides are distinctive in their repetitive sequences, -Gly-X-Y- with a preponderance of Pro as X or Y in which many Pro and Lys at Y are hydroxylated. These hydroxyPro and hydroxyLys form hydrogen bonds to stabilized the triplet helix.

A large number of the structural proteins involved in maintaining cell shape, organizing cytoplasm and in movement are coils of two or three α-helices wound around each other to form a left-handed superhelix. Many of coiled-coil proteins have segments of polypeptide chain with different conformations at one or both ends. For example, myosin of muscle is a long coil that has globular heads at each of the two-carboxyl ends, which carry out the enzymatic and movement functions.

Silk fibroin is the structural protein of spider webs and silkworm cocoons. The silkworm protein consists of antiparallel β-sheets with ~50 repeated Gly, Ala and Ser rich sequences interspersed with irregular regions. Collagen is the main constituent of the bones, tendons, skin, ligaments, blood vessels and supporting membranous tissues. Collagen is composed of three parallel polyproline type II helices, each formed by Pro and hydroxyPro (Hyp)-rich Gly-X-Y repeats. Collagen polypeptide chains typically have just over 1000 residues producing a triple helix 14 Å in diameter and 3000 Å in length. Collagen functions by aggregating side-by-side into microfibrils, which assemble into large supramolecular arrays.

5.6.6 Circular (cyclic) proteins

Conventional proteins are linear polypeptide chains of amino acids whose sequences are encoded by DNA and fold into 3D conformations that define their biological functions.

Conceptually, the amino and carboxyl termini of a polypeptide chain are flexible and amendable to form a peptide bond. The formation of the terminally linked peptide bond yields circular (cyclic) proteins with circular backbones. Cyclic peptides such as cyclosporin are known. These peptides tend to be less than 12 amino acids in size, contain modified amino acids and are generally metabolic products. Whereas circular proteins are 14–70 amino acids in size, true gene products (encoded by DNA) with well-defined 3D structures. They occur in microorganisms, plants and animals, as products for an enhanced stability or involvement in host defense (Trabi and Craik, 2002). Several naturally occurring circular proteins are listed in Table 5.10. CyBase (http://research.imb.uq.edu.au/cybase) is the curated database for cyclic proteins.

5.6.7 Representation of protein topology

Protein topology cartoons (TOPS) are two-dimensional schematic representations of protein structures as a sequence of secondary structure elements in space and direction (Sternberg and Thornton, 1977; Michalopoulos *et al.*, 2004). TOPS are compact and highly abstract descriptions, reducing the protein fold to a sequence of secondary structure elements (SSEs) and three sets of pair-wise relationships between them, i.e. hydrogen bonds relating parallel and antiparallel β strands, spatial adjacencies relating neighboring SSEs, and chirality of selected supersecondary structures. TOPS of trypsin domains, as shown in Figure 5.9, have the following symbolism:

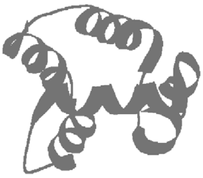
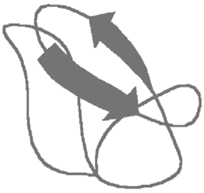
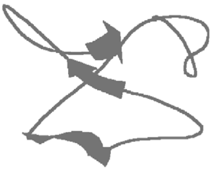
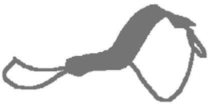
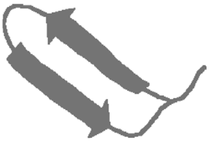
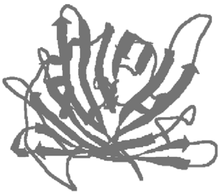
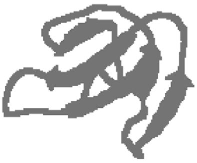
- Circular symbols represent helices (α and 3_{10}).
- Triangular symbols represent β strands
- The peptide chain is divided into a number of fragments and each fragment lies in only one domain.
- Each fragment is labeled with an integer (i), beginning at N_i and ending at C_{i+1}, with the first fragment being $N_1 \rightarrow C_2$.
- If the chain crosses between domains, it leaves the first at C_{i+1} to join the next N_{i+2}.
- Each secondary structure element has a direction (N to C), that is either up (out of the plane of the diagram) or down (into the plane of the diagram).
- The direction is up if the N terminal connection is drawn to the edge of the symbol and the C terminal connection is drawn to the center of the symbol. Otherwise the direction is down if the N terminal connection is drawn to the center of the symbol and the C terminal is drawn to the edge.
- For β strands, up strands are indicated by upward pointing triangles, whereas down strands are indicated by downward pointing triangles.

The topology cartoons can be browsed and searched at TOPS server (http://www.tops.leeds.ac.uk/).

5.6.8 Accessible surface of folded structures

The contact surface of a protein molecule is designated as those parts of the van der Waals surface that comes in contact with a rolling chemical probe, typically a water molecule (approximated as a spherical probe). When the water molecule is simultaneously in contact with more than one protein atom, its border surface defines the re-entrant surface, which is the area described by the center of a spherical water molecule of radius R_w (taken as 1.4 Å). The contact surface and the re-entrant surface combined trace a continuous surface known as the (molecular) accessible surface. The accessible surface area (A_s) in $Å^2$ of a

TABLE 5.10 Some naturally occurring circular proteins

Circular protein	Source	#Aa and pdb file	Structure	Characteristics/features
Bactericin AS-48	*Enterococcus faecalis*	70 1O84		Antibacterial protein consists of five α helices connected by short turns that enclose a hydrophobic core.
Kalata B1/B2	Tropical plant, *Rubiaceae* family	~30 1PT4		Macrocyclic polypeptides with ~30 amino acids with six conserved Cys that form three knotted disulfide bonds with uterotonic activity.
McoTI-II	Seeds of vine, *Momordica cochinchinensis*	34 1IB9		Six conserved Cys (3 knotted disulfides) with trypsin inhibitory activity. Small triple stranded β sheet is associated with cyclic cysteine knot motif.
SFTI-1	Seeds of sunflower, *Helianthus annuus*	14 1SF1		This potent trypsin inhibitor is structurally similar to legume Ser protease inhibitors and is stabilized by two β strands and a disulfide bond.
RTD-1	Leukocytes of *rhesus macaquesw*	18 1HVZ		Antibacterial rhesus theta defensin-1 with 6 Cys (3 ladder-like disulfides) and 5 Arg consists of two β strands connected by two tight turns. It posses antibacterial activity.
Green fluorescent protein	Jelly fish, *Aequorea victoria*	245 1KP5		Dimeric luminescent cyclic protein with antiparallel β-pleated structure.
Fasciculin 2	*Dendroaspis antgusticeps*	61 1KU6b		Venom toxin (three-loop toxin) which is a potent inhibitor of mammalian acetylcholinesterase. The cyclic protein consists of 8 Cys and six antiparallel β strands.

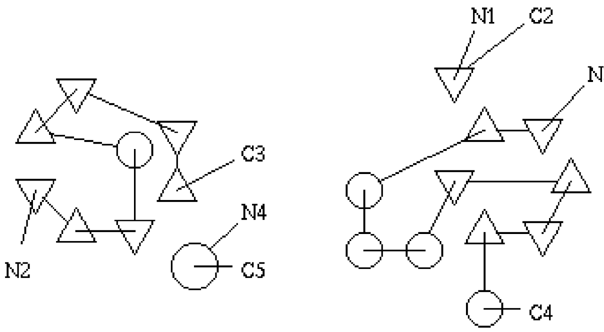


Figure 5.9 TOPS diagrams for trypsin domains

small monomer protein (molecular weights ranging between 4–35 kDa) can be estimated from its molecular weight, M (Miller *et al.*, 1987a) by:

$$A_s = 6.3\,M^{0.73}$$

This is only 34 ± 11% of the surface area of the unfolded polypeptide chain. Since the accessible surface area of proteins in an extended conformation is proportional to their molecular weight by

$$A_t = 1.45\,M$$

it follows that the potential surface buried by the protein folding, A_b can be estimated according to

$$A_b = 1.45\,M - 6.3\,M^{0.73}$$

for small monomeric proteins.

The volume (V_m) of a typical monomer protein can be estimated by

$$V_m = 1.27\,M\,\text{Å}^3/\text{dalton}$$

The accessible surface area of an oligomeric protein is related to its molecular weight (Miller *et al.*, 1987b) by

$$A_s = 5.3\,M^{0.76}$$

The interiors of proteins are densely packed with adjacent atoms, generally in van der Waals contact.

Despite a growing focus on proteins whose folding is assisted in a kinetic sense by accessory proteins called chaperonins, there remains general agreement that equilibrium for 3D structures of proteins are determined by their amino acid sequences. There is a good reason to believe that the chemistry of amino acid residues does affect the overall folding of protein structures. X-ray crystallographical analysis demonstrates that hydrophilic and ionizable residues are located mainly on the surface and hydrophobic residues are mainly buried in the interior of the protein molecules. This distribution of amino acid residues of proteins is known as the ‘nonpolar-in, polar-out’ rule. Virtually all ionized groups of soluble proteins are exposed to the solvent. There is a tendency for oppositely charged groups to be near each other on the surface where they can form salt bridges in solution. Ionized pairs of acidic and basic groups rarely ever occur in the interiors of proteins. Hydrophobic side chains tend to cluster together via hydrophobic interactions. These residues are mainly involved in the formation of the secondary structures. The protein interiors are occupied predominately by nonpolar residues. Small proteins have few completely buried residues and only ~15% of the residues are totally inaccessible to solvent in large proteins. The residue is considered buried if more than 95% of its surface area is inaccessible to the solvent.

5.7 CLASSIFICATION OF PROTEIN STRUCTURES

Proteins often display substantial similarity in sequence and 3D structure, since many are derived from a basic complement of autonomously folding units (domains). This allows us to group proteins into a hierarchy of families, superfamilies and folds. Domains within protein structures are defined as spatially distinct structures that could conceivably fold and function in isolation. The concept of fold thus allows grouping of related structures for descriptive purposes and simplifies problems related to the encoding of structure in primary sequences.

A classification system, such as Structural Classification of Proteins (SCOP) (Lesk and Chothia, 1984; Anreeva *et al.*, 2004) categorizes structure domains based on secondary structural elements within a protein into α structure (made up primarily from α helices), β structure (made up primarily from β strands), α/β structure (comprised of primarily β strands alternating with α helices), α + β structure (comprised of a mixture of isolated α helices and β strands) and others. In this classification (Brenner *et al.*, 1996), only the core of the domain is considered. Therefore it is possible for an all-α structure to have a very small amount of β strand outside the α-helical core. Similarly an all-β protein may have a small presence of α or 3_{10} helix. The SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/) can be summarized as followings:

5.7.1 α-Helical proteins

Most proteins contain some helix. In a significant number of proteins, the secondary structure is limited to α-helices. For example:

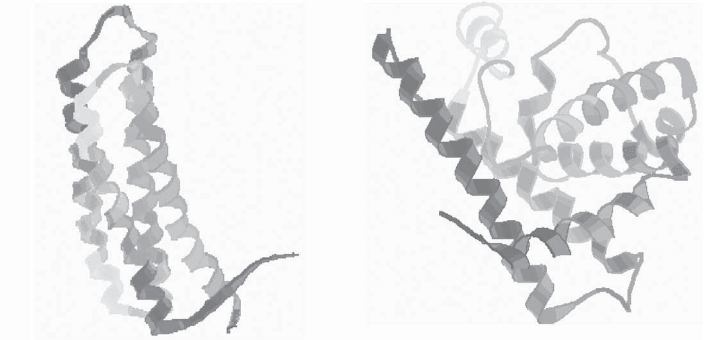
- There is a class of small proteins (haemerythrin, repressor of primer, Figure 5.10A) that have the form of a helix bundle in up-down superfold.
- Globins such as myoglobin (Figure 5.10B), hemoglobin are rich in helices in globin superfold.
- The cytochromes C are another family of haem proteins with exclusively helical structures.
- Very large proteins (e.g. citrate synthase) can also have a purely helical secondary structure.
- Membrane attached receptor proteins (e.g. bacteriorhodopsin) contain transmembrane α-helices.

For individual pairs of interacting helices, certain regularities in the patterns of tertiary-structural interactions between pairs of helices (e.g. orthogonal or parallel arrangements) have been observed.

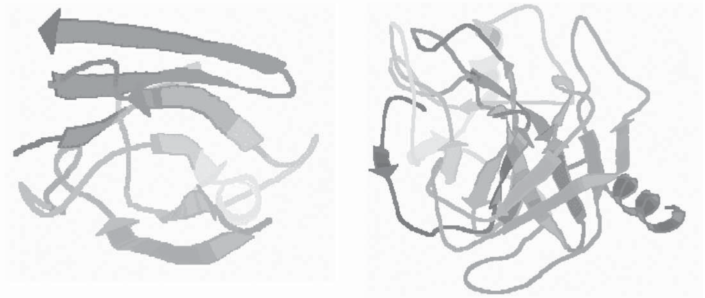
5.7.2 β-Sheet proteins

Domains in which the secondary structure is almost exclusively β-sheets tend to contain two sheets packed face to face. There are two major classes: those in which the strands are almost parallel and those in which the strands are almost perpendicular. The strands

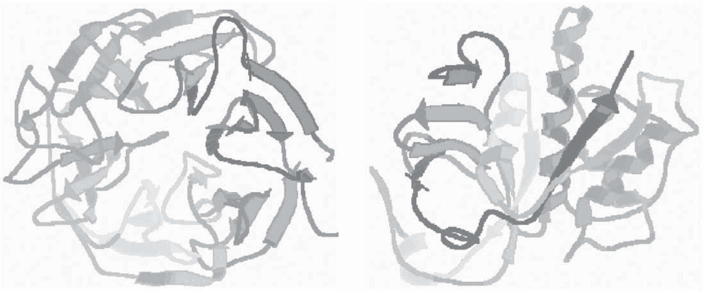
(a) Primer repressor (1RPR.pdb) (b) Myoglobin (1MBO.pdb)



(c) Transthyretin (1BMZ.pdb) (d) Trypsin (1QQU.pdb)



(e) Neuraminidase (1EUR.pdb) (f) Papain (9PAP.pdb)



(g) Lactate dehydrogenase (2LDB.pdb) (h) Triosephosphate isomerase (1TIM.pdb)

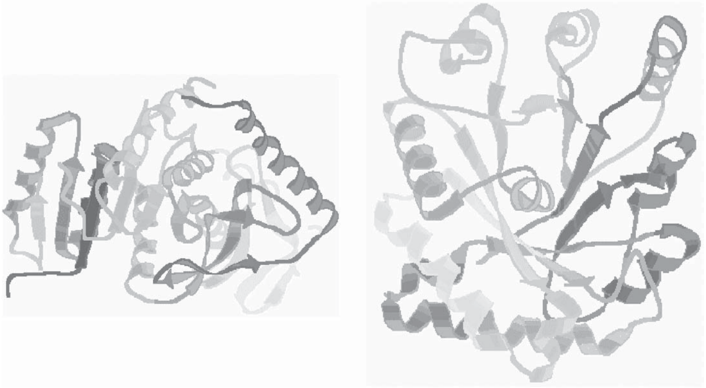


Figure 5.10 Representative protein domain structures
Representative protein structures according to SCOP classification are illustrated with spectral color by relative N to C position. They are: (*a*) repressor of primer with parallel up-down helical pairs (*b*) sperm whale myoglobin for α-domain with orthogonal helical pairs in globin superfold (*c*) human transthyretin (prealbumin) for β-domain with parallel sheets, (*d*) porcine β-trypsin for β-domain with orthogonal sheets, (*e*) *Micromonospora viridifaciens* neuraminidase for β-domain with propeller-like structure, (*f*) papain of papaya latex for α + β domain, [*g*] *Bacilus stearothermophilus* lactate dehydrogenase for α/β with open β-α-β motif and (*h*) chicken muscle triosephosphate isomerase for α/β domain with close β-barrel.

of a sheet can vary in polarity (parallel and antiparallel) and in connectivity, giving great topological variety.

5.7.2.1 Parallel β-sheet proteins.

The most common arrangement in this class contains four strands for each sheet to form the natural twist. However, the directions of the strands are not all equivalent and the connectivities of the strands are different, such as prealbumin (e.g. transthyretin, Figure 5.10C) in a Greek key topology.

5.7.2.2 Orthogonal β-sheet proteins.

An alternative way of packing two β-sheets together is with the strands in the two sheets almost perpendicular in a trefoil superfold. Each domain of the serine proteases (e.g. trypsin, Figure 5.10D) shows this arrangement.

5.7.2.3 Other β-sheet proteins.

Influenza neuraminidase contains an unusual β-sheet propeller-like structure (Figure 5.10E). Ascorbate oxidase and galactose oxidase are large β-sheet proteins that contain parallel β-sheet domains in a jellyroll- or sandwich-like structure.

5.7.3 α + β proteins

Many proteins, such as lysozyme (Figure 5.1), contain both α-helices and β-sheets, but do not have the special structures created by alternating β-α-β patterns. In the sulfhydryl proteases, such as papain (Figure 5.10F) and actinidin, the strands of sheets and the helices tend to be segregated in different regions of space.

5.7.4 α/β proteins

The supersecondary structure consisting of a β-α-β unit with the hydrogen bonded parallel β-strands forms the basis of many enzymes, especially those that bind nucleotides or related molecules. The strands form a parallel β-sheet. In some cases, there is a linear β-α-β-α-β- . . . arrangement, but in other cases the β-sheet closes on itself with the last strand hydrogen-bonded to the first.

5.7.4.1 Linear or open β-α-β proteins.

Many proteins that bind nucleotides contain a domain made up of six β-α units with a special topology. The long loop between β_C and β_D tends to create a natural pocket for the nucleotide ligand. The N—H groups in the last turn of the α_A helix are well positioned to form hydrogen bonds to phosphate oxygen atoms of the ligand. The NAD-binding domains of horse liver alcohol dehydrogenase and lactate dehydrogenase (Figure 5.10G) are typical. Other dehydrogenases have similar domains. Flavodoxin and adenylate kinase contain a variation on the theme. They have five strands instead of six. Dihydrofolate reductase has eight strands.

5.7.4.2 Closed β-α-β barrel structures.

Chicken triose phosphate isomerase (Figure 5.10H) is typical of a large number of structures that contain eight β-α-units, in which the strands form a sheet wrapped around into a closed structure, cylindrical in topology. The helices are on the outside of the sheet.

5.7.5 Multidomain structures

Proteins may have multiple domains but the different domains of these proteins have never been seen independently of each other, therefore accurate determination of their boundaries is not possible and perhaps not meaningful.

5.7.6 Membrane and cell surface proteins

Within membrane and cell surface proteins, the membrane proteins are defined based on the number of helices that span the membrane.

5.7.7 Irregular and small proteins

There are structures that contain very few of their residues in helices and sheets. These tend to be stabilized by disulfide bridges or by metal ligands. For instance:

a) In the case of wheat germ agglutinin, there are numerous disulfide bridges.

b) In the case of ferridoxin, there are ion-sulfur clusters.

c) The kringle structure occurring in many proteins contains disulfide bridges as well as several short stretches of two-stranded β-sheet.

The CATH protein domain database (http://www.biochem.ucl.ac.uk/bsm/cath) is a hierarchical classification of protein domain structures into evolutionary families and structural groupings, depending on sequence and structure similarity (Pearl *et al.*, 2005). The protein domains are clustered and classified according to four major (original) levels.

Level 1. *Class* (C): Based on the secondary structure content of the given domain, three major classes, namely mainly-α, mainly-β, and both α and β are recognized, plus class 4 for few secondary structures,

Level 2. *Architecture* (A): Description of the overall shape of the domain structure as determined by the orientations of the secondary structures in 3D space independent of their connectivity.

Level 3. *Topology* or *Fold* (T): Protein domains with significant structural similarity but no sequence or functional similarity. They have similar number and arrangement of secondary structures and similar connectivity.

Level 4. *Homologous superfamily* (H): Proteins domains thought to share a common evolutionary ancestor and can be therefore be described as homologous.

Thus the architecture refers to the overall shape of the protein, whereas the topology or fold group describes the relative orientations of the secondary structures in 3D and the order in which they are connected. The Class level is derived automatically, while the architecture is assigned manually. The assignments of structures to topology families and homologous superfamilies are made by sequence and structure comparisons. Four levels are added in the hierarchical matches. They are:

1. *Sequence family* (*S*): Domains that have a sequence identity of greater or equal to 35% because domains that share 35% or more identical residues nearly always have highly similar structures.
2. *Non-identical* (*N*): Domains that have a sequence of identity of greater or equal to 95% having almost identical structures so this level is often used to provide non-redundant datasets.
3. *Identical* (*I*): Identical domains that share 100% sequence identity derived from PDB.
4. *Domain* (*D*): Semi-independent folding unit.

Only protein structures solved to resolution better than 3.0 Å from the Protein Data Bank are considered. The 3D templates are generated with Conserved Residue Attributes

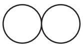
(CORA) for recognition of structural relatives in each fold group (Orengo, 1999). The Class, Architecture, Topology, Homology (CATH) Architectural descriptions that denote the arrangements of secondary structures are given in Table 5.11.

5.8 QUATERNARY (SUBUNIT) STRUCTURES OF PROTEINS

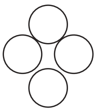
Quaternary structure describes the formation of oligomeric proteins from monomeric subunits (polypeptide chains). The majority of globular proteins possessing subunit structures consist of two or four subunits. Few proteins have more than 12 subunits. Moreover, the majority of oligomeric proteins have an even number of subunits. These subunits can function either independently of each other or cooperatively so that the function of one subunit is dependent on the functional state of other subunits. Each polypeptide chain subunit is folded into an independent globular conformation, which then interacts with other monomer (protomer). Two fundamental types of interaction between identical monomers are possible. Isologous association involves the same surfaces on monomers, which associate to produce a symmetric oligomer. Heterologous association involves different sites from nonequivalent monomers that are complementary and largely nonoverlapping in the oligomeric interactions.

Subunits of an oligomeric protein associate with one another in regular patterns, which are succinctly, by the nomenclature, provided with a particular type of symmetry. Any system that is stabilized by multiple weak interactions, such as subunit interactions in the quaternary structures, will possess some type of well-defined symmetry since this ensures the formation of the maximum number of most stable bonds. A structure is said to possess symmetry if it is possible to operate on in such a way that the transposed structure is indistinguishable from its original state. The operation of transposing one part to the position of a symmetrically related part in the structure is termed a symmetry operation, including rotation, reflection and inversion. Since proteins are chiral, reflection and inversion symmetries are not possible because they produce the enantiomorphic forms. Consequently, the symmetry element (the geometric entity, e.g. point, line or plane with which a symmetry operation is performed) is restricted solely to rotation around an axis passing a point (the geometric center), which remains unchanged in the operation. Symmetry of this kind is called point symmetry possessing point groups. The quaternary structure of oligomeric proteins is considered as the structure with point symmetry possessing protomers as point groups. Two important point symmetries for oligomeric globular proteins are cyclic and dihedral symmetries:

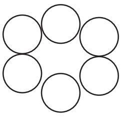
Cyclic (Rotation) symmetry: It is the simplest arrangement possible, involving isologous association of subunits. The disposition of the subunits is such that there is an n-fold rotational axis where n is the number of subunits. A rotation of 360°/n transposes the structure into itself. This type of symmetry is designated as C_n.



C_2: Linear



C_4: Square



C_6: Hexagon

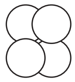
Alcohol dehydrogenase is a dimeric protein with C_2, and pyruvate carboxylase is a tetrameric protein with C_4 symmetries respectively. The cyclic symmetry allows the polar

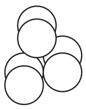
TABLE 5.11 Top two levels of CATH classification

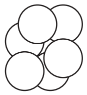
Architecture		
Code No.	Level name (Description)	Example
Class 1, Mainly Alpha:		
10	Orthogonal bundle	Aldehyde dehydrogenase, domain 2
20	Up-down bundle	ATP Synthase, domain 2
25	Horseshoe	70 kDa Soluble lytic transglycosylase, domain 1
40	Alpha solenoid	Peridinine chlorophyll protein, chain M
50	Alpha/alpha barrel	Glycosyl hydrolase, domain 2
Class 2, Mainly Beta:		
10	Ribbon	Trypsin inhibitor
20	Single sheet	Rubrerythrin, domain 2
30	Roll	Phosphomannose isomerase, domain 1
40	Barrel	Endoglucanase V
50	Clam	Bacteriochlorophyll-a protein
60	Sandwich	Galactose oxidase, domain 1
70	Distorted sandwich	Topoisomerase I, domain 3
80	Trefoil	Acidic Fibroblase growth factor, subunit A
90	Orthogonal prism	Agglutinin, subunit A
100	Aligned prism	Vitelline membrane outer layer protein I, subunit A
102	3 Layer sandwich	Rieske iron-sulfur protein
110	4 Propeller	Hemopexim
115	5 Propeller	
120	6 Propeller	Neuraminidase
130	7 Propeller	Methylamine dehydrogenase, chain H
140	8 Propeller	Methanol dehydrogenase
150	2 Solenoid	Alkaline protease, subunit P, domain 1
160	3 Solenoid	UDP-NAc-glucosamine acyltransferase, domain 1
170	Complex	Phosphoenol pyruvate carboxykinase, domain 1
Class 3, Alpha and beta:		
10	Roll	Elastase, domain 1
15	Super roll	Bactericidal permeability incr. Protein, domain 1
20	Barrel	Aconitase, domain 4
30	2 Layer sandwich	α-Amylase, domain 2
40	3 Layer (aba) sandwich	Lysozyme
50	3 Layer (bba) sandwich	Restriction endonuclease, domain 2
55	Structural genomics	
60	4 Layer sandwich	Deoxyribonuclease I, subunit A
65	Alpha-beta prism	UDP-NAG-Cbov. transferase, chain A, domain 1
70	Box	Proliferating cell nuclear antigen
75	5 Stranded propeller	L-Arg/Gly aminotransferase, chain A
80	Horseshoe	Ribonuclease inhibitor
85		Sulfite reductase hemoprotein, domain 2
90	Complex	Glutamine synthetase, domain 1
100	Ribosome	
Class 4, Few secondary structures:		
10	Irregular	Glucose oxidase, domain 2

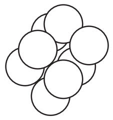
structures of oligomeric membrane protein subunits to persist in their quaternary structures. Membrane-bound proteins function at the boundary between two different environments. The membrane itself is nonpolar while the exterior environment is aqueous. They consequently require a polar structure that projects two different surfaces, one toward each environment. Cyclic symmetry matches this requirement. Examples are Influenza meuraminidase (Colman *et al.*, 1983) F1 ATPase (Abrahams *et al.*, 1994).

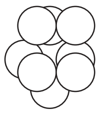
Dihedral symmetry: Oligomeric soluble proteins usually associate with dihedral symmetry designated as D_n, with the number of subunits being 2n. Their dihedral symmetry is limited to an oligomer containing an even number of subunits, which associate via heterologous interactions. This symmetry is present if n two-fold axes exist at right-angles to any single n-fold axis.


D_2:
Tetrahedron


D_3:
Trigonal prism


D_3:
Octahedron


D_4:
Cube


D_4:
Square antiprism

The classic oligomeric protein, hemoglobin, is a dimer of dimers (Liddington *et al.*, 1992). Other examples are octomeric mandelate recemase (Neidhart *et al.*, 1991) and $\alpha_6\beta_6$ dodecameric aspartate carbamoyl transferase (Stevens *et al.*, 1990).

Other symmetries for polymeric proteins include:

Icosohedral symmetry: This point symmetry may not be of great importance for the majority of proteins, because only structures with 12n subunits may belong to these point groups. The icosohedral 5 : 3 : 2 point group generates 60 asymmetric units, the largest possible number for a point group. It has 12 (=60/5) five-fold axes, 20 (=60/3) three-fold axes, and 30 (=60/2) two-fold axes. All spherical viruses known have approximately this symmetry. The augmented icosahedral shells of viruses should contain $60T$ subunits, 60 of which obey the icosahedral symmetry (Caspar and Klug, 1962). Examples are satellite tabacco necrosis virus, $T = 1$ virus (Jones and Liljas, 1984) and tomato bushy stunt virus, a $T = 3$ virus (Harrison *et al.*, 1978).

Helical symmetry: The polymeric proteins of filamentous viruses and the cytoskelton possess helical symmetry, in which subunits are related by a translation, as well as a rotational component. Actin, myosin, tubulin and various other fibrous proteins all interact with helical symmetry, which is often called ‘screw’ symmetry. Screw symmetry, which relates the positions of adjacent subunits, combines a translation along the helix axis with the rotation. Actin forms a two-stranded helix of globular actin subunits. However, important variations in the helix parameters occur (Egelman *et al.*, 1982). The rise per subunit is relatively constant, but the twist or relative rotation around the helix axis is highly variable. This polymorphic tendency is probably important for the smooth functioning of muscle contraction, which involves considerable force generation.

Different types of symmetry are appropriate for proteins localized in different environments. In assemblies with intermediate numbers of protomers, the symmetry is often reduced by the inclusion of genetic variants, in order to restrict the tendency to polymerize. Finally, symmetry tends to be preserved when subunits change conformation, giving rise to the phenomenon of cooperativity, which is often crucial to function.

5.9 QUINTERNARY STRUCTURE EXEMPLIFIED: NUCLEOPROTEINS

Nucleoproteins are at the core of cellular processes utilizing genetic information. Most proteins in the nucleoprotein assemblies are basic and small, having molecular weights of 10–25 kDa. Many of these proteins associate with a single nucleic acid and generally with a single stretch of nucleic acid helix.

5.9.1 Chromosomes

Eukaryotic chromosomes consist of a complex of DNA, RNA and protein called chromatin. The individual chromosomes assume their familiar condensed forms only during the interphase between cell divisions. The genetic materials of a typical human cell (20 μm in diameter) consists of 23 pairs of dsDNA ($\sim 3 \times 10^9$ bp) with an average of 0.65×10^8 bp (3×10^9 bp/46 chromosomes) corresponding to a DNA molecule of approximately 2.2 cm long (0.65×10^8 bp $\times 3.4 \times 10^{-8}$ cm/bp) per chromosome. This indicates that the DNA molecules of chromosomes of a meter long ($46 \times 2.2 \times 10^{-2}$ m) must be packed into a nucleus perhaps 5 μm long. This high degree of condensation (packing) results from three levels of folding in the quinternary structural organization of the DNA-protein complex.

(1) *Nucleosome*: There are two classes of chromosomal proteins; histones (structural proteins) and nonhistone chromosomal proteins (mainly regulatory proteins). Histone Database (http://research.nhgri.nih.gov/histones/) compiles histone folds and sequences. The histones are relatively small, positively charged Arg or Lys rich proteins that interact via ionic interaction with the negatively charged polynucleotide backbone of DNA. Five distinct histones are known, H1, H2A, H2B, H3 and H4. The DNA duplex is wrapped around the histone octomer, consisting of $(H2A)_2(H2B)_2(H3)_2(H4)_2$ in a left-handed superhelix, to form nucleosomes (nucleosome core particles) and one copy of the linker histone, H1. All four histones contain a common fold in which a long central helix is flanked on each side by a loop and a shorter helix. The histone octamer has a twofold symmetry structure comprised of two H2A–H2B dimers flanking centrally located $(H3–H4)_2$ tetramer. The surface of the octamer forms a left-handed helical ramp (1.65 turns) whose helix axis is perpendicular to the twofold axis of the octamer. The DNA is apparently bound along this ramp, which lined by numerous positively charged Arg and Lys side chains (Arents and Moudrianakis, 1991). The fifth histone (H1) probably binds to the two ends and middle of the superhelical (two turns) DNA of nucleosomes, which are then packed in a helical filament (10-nm nucleosomal filament). The 10-nm nucleosomal filament formed at low ionic strength represents the first level of compaction.

(2) *30-nm Filaments*: At physiological ionic strengths at an increased salt concentration, the H1-containing nucleosomal filament folds into a zigzag conformation and then forms a 30-nm-thick filament. The filament is probably constructed by winding the 10-nm filament into a solenoid with ~6 nucleosomes per turn and a pitch of 11 nm (diameter of the nucleosome). The solenoid is stabilized by the H1 molecules whose extended N-terminal and C-terminal arms are thought to contact adjacent nucleosomes by interacting with neighboring H1 in a head-to-tail fashion. This seals the ends of the DNA turns to the nucleosome core and to organize the DNA linkers of consecutive nucleosomes (Widom and Klug, 1985).

(3) *Radial loops*: The 30-nm filament then form long DNA loops of variable length (15–30 μm in lengths corresponding to 45–90 kb). These loops are then arranged radially about the circumference of a single turn to form a miniband unit of the chromosome.

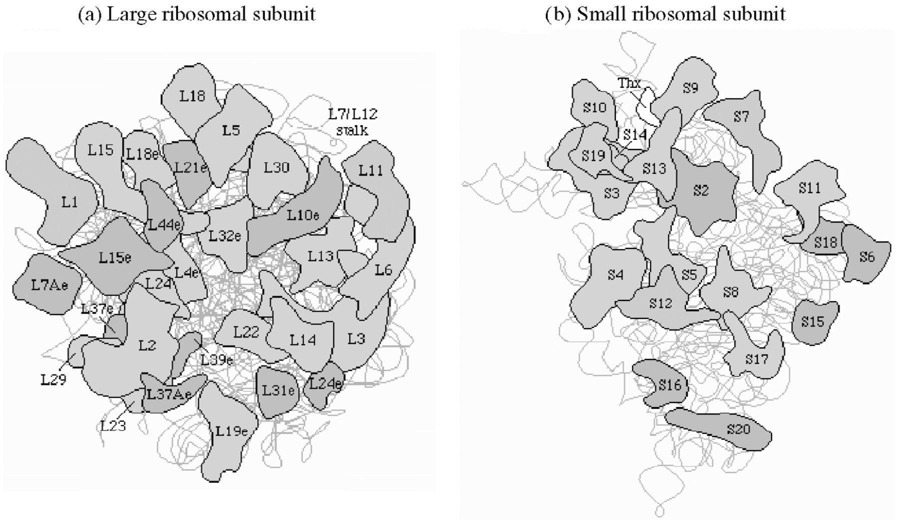
5.9.2 Ribosomes

Ribosomes are compact ribonucleoprotein particles found in the cytosol of all cells. It is a spheroidal particle that can be dissociated into two unequal subunits. The small subunit is a roughly mitten-shaped particle whereas the large subunit is arm-chair like with three protuberances on one side. The small subunit is primarily associated with the control of mRNA binding, decoding and fidelity, whereas the large subunit is attributed to bind tRNA and to mediate peptidyl transfer, though most biological activities of ribosomes occur at the subunit interface. Each subunit consists of unequal size and number of polyribonucleotide (rRNA) and polypeptide chains. Functionally, rRNAs are probably involved in mRNA selection, frame-shift suppression, ribosomal subunit association, tRNA binding and various translational processes, while ribosomal proteins are important for their effects on folding and fine-tuning of the conformation of rRNA during ribosome assembly. Consistent with this, the majority of proteins appear to have multiple RNA-binding sites and probably interact with several regions of rRNA. Furthermore, many of the binding sites in rRNA are characterized by single-stranded loops and bulges that can facilitate these interactions. Although eukaryotic and prokaryotic ribosomes resemble each other in both structure and function, they differ in other details such as the molecular size and components (Table 5.12).

Ribosomal RNA (rRNA) molecules fold extensively into characteristic secondary structures as a consequence of intramolecular hydrogen bonding. Sequence comparisons of the corresponding rRNAs from various species indicated that they maintain evolutionarily conserved secondary structures rather than their base sequences. Ribosomal proteins from the small and large subunits are designated with the prefixes *S* and *L* respectively, followed by a number indicating their position, from upper left to lower right, on a 2D gel electrophoreogram (corresponding roughly to the order of decreasing molecular mass). The distributions of prokaryotic ribosomal proteins in large and small ribosomal subunits are depicted in Figure 5.11. The ribosomal proteins are of a wide variety of structural types, and many of the common folding domains such as α/β, α-helical bundle and β-ribbon are found (Ramakrishnan and White, 1998). Many of the known 3D structures of ribosomal proteins display homologous structural motifs that consist of a 3-stranded antiparallel β sheet with the latter two strands connected by an α helix (Leijonmarck *et al.*, 1988) designated as the RNA-recognition motif (RRM).

TABLE 5.12 Components of prokaryotic and eukaryotic ribosomes

	Prokaryote (*E. coli*)		Eukaryote (Rat liver)	
Ribosome (kDa)	70S (2.52×10^3)		80S (4.22×10^3)	
Subunits (kDa)	30S (0.93×10^3)	50S (1.59×10^3)	40S (1.40×10^3)	60S (2.82×10^3)
RNA (#Nucleotides)	16S RNA(1542)	23S (2904)	18S (1874)	28S (4718)
		5S (120)		5.85S (160)
				5S (120)
Protein (kDa)	857		1700	
	370	487	700	1000
Polypeptides	21	31	33	49


(a) Large ribosomal subunit

(b) Small ribosomal subunit

Figure 5.11 Distribution of ribosomal proteins
Sketches of large subunit (*Haloarcula martsmortsui*) and small subunit (*Thermus aquaticus*) of ribosome particles showing distribution of ribosomal proteins are retrieved from KEGG at http://www.genome.jp/kegg/pathway.html.

5.9.3 Spliceosome and splicing activities

Spliceosome is a functional complex, in which splicing small nuclear ribonucleoprotein (snRNP) is complexed with the pre-mRNA (Krämer, 1996). Fully assembled spliceosome sediments have 50–60 S. Major components in the alignment of the reactive sites and the formation of the catalytic core of the spliceosome are the U1, U2, U4/U6 (base-paired extensively) and U5 small nuclear RNAs (snRNAs), which are associated with several proteins (Sm proteins) common to all snRNAs to form snRNPs. These proteins bind to the conserved sequence RAU$_{3-6}$GR present in all snRNAs, except U6. In addition, all spliceosome snRNAs contain unique proteins, some of which play essential roles in splicing. The 5′ end of U1 snRNA is complementary to the conserved sequence at the 5′ splice site. Interactions of U2, U5 and U6 snRNAs with the pre-mRNA and with one another are crucial elements of the catalytic core of the spliceosome. U1 and U4 snRNAs do not appear to serve an essential function during catalysis. But the former (U1) is important for the initial recognition of the pre-mRNA and the latter (U4 by base-pairing to U6) acts as a regulatory element of U6 by sequestering the sequences that participate in the formation of the active site. U6 snRNA is probably involved in the direct catalysis.

Major protein components participating in the splicing activities include the SR family of splicing proteins, pyrimidine tract-binding proteins (PTB) and branch site-binding proteins. The SR proteins represent a family of splicing factors containing a domain rich in Ser and Arg, therefore termed SR proteins and are designated as SRPn (where n is apparent molecular mass in kDa such as SRp20, SRp40, SRp55 etc.). The N-terminal part of the SR protein consists of an RNA recognition motif (RRM, also known as RNA-binding domain) and a limited but significantly homologous domain (Ψ-RRM). The RRM is organized into four antiparallel β sheets and two α helices, which are arranged in the order β1-α1-β2-β3-α2-β4. The Ψ-RRM has the characteristic sequence SWQDLKD. The two domains are usually separated by a glycine-rich hinge. The C-terminal part of the SR protein harbors a domain enriched in Ser and Arg (SR domain), many of which repeat as RS or SR dipeptides. The variable length of this domain contributes to the size difference observed among individual SR proteins. The SR domain is highly phosphorylated. Thus the activity of SR proteins can be modulated by phosphorylation/dephosphorylation. By altering the ratio of positive and negative charges, phosphorylation/dephosphorylation is likely to affect protein folding, RNA binding activity, protein–protein interactions of SR proteins and, consequently spliceosome assembly and catalysis. SR proteins interact with pre-mRNA and enhance the binding of U1 snRNA to 5′ splice sites. Individual SR proteins differ in their recognition of specific RNA sequences and therefore the ability to commit various pre-mRNA to the splice pathway, as well as in their interactions with competing 5′ splices and exonic splicing enhancers. SR proteins appear to be involved in the alternative splice site selection. Thus SR proteins with specific RNA-binding properties are likely to affect the maturation of a wide range of pre-mRNA substrates in constitutive and alternative splicing. Both RRM and Ψ-RRM are essential for constitutive splicing and alternative splice-site switching.

Most introns of higher eukaryotes contain a region of high pyrimidine content located between the branch site and the conserved AG dinucleotide at the 3′ splice site. The PTB plays an essential role in the definition of the 3′ splice site at the earliest stage of spliceosome assembly. PTB is a protein of ~62 kDa that show a preference for binding to polypyrimidine tract-containing RNAs in the formation of pre-splicing complex. In addition to pyrimidine-rich RNA, PTB also binds single-stranded pyrimidine-rich DNA. Different proteins are involved in the recognition of the branch site at each step of spliceosome assembly and catalysis. The dynamic nature of these interactions parallels the conformational rearrangements in the spliceosome owing to changes in the base-pairing interactions of pre-mRNA and snRNAs.

5.10 CONFORMATIONAL ENERGETICS

Both covalent and noncovalent bonds/interactions contribute to maintaining the 3D structures of proteins. Disulfide bonds are the most important covalent bonds that appear to be more prevalent among small or extracellular proteins with single polypeptide chain. The formation of disulfide bonds from sulfhydryl groups of Cys is very specific. The bonds confer added stability by complementing the noncovalent interactions that determine the energetically favorable conformation of proteins. The folded conformation of natural proteins is generally stabilized by ~17–42 kJ/mol. Among the noncovelent interactions are hydrogen bonds, hydrophobic interactions, electrostatic interactions (ionic bonds) and dispersion forces.

To evaluate the total free energy change, ΔG for the formation of a native structure form the random polypeptide chain, i.e. $\Delta G = -RT \ln K = \Delta H - T\Delta S$ for the process:

$$\text{Random polypeptide chain} \overset{K}{\leftrightarrow} \text{Native protein}$$

The free energy change of the contributing forces/bondings can be expressed by:

$$\Delta G = \Delta G_{conf} + \Sigma\Delta g_{i,int} + \Sigma\Delta g_{i,s} + \Delta W_{ei} + \Delta G_{s\text{-}s}$$

where

ΔG_{conf}: Conformational free energy change, which opposes the transition from the multiplicity of random conformations to the single native state due to unfavorable entropic restriction ($-\Delta S_{conf}$). It is estimated as $\Delta G_{conf} \approx +10$ kJ/mole residue at 27°C (310 K) ($-\Delta S_{conf} \approx 46$ J/deg/mole residue and $\Delta H_{conf} \approx -4.2$ kJ/mole residue).

$\Delta g_{i,int}$: Free energy changes for short range interactions, i.e. hydrogen bonding, hydrophobic interactions, dispersion forces.

TABLE 5.13 Estimated contribution of individual factors to stability of a hypothetical protein of 100 residues at 25°C

Factor	Free energy difference between N and D (10^3 kJ/mol)
Conformational entropy	−1.255—−4.184
Unfavorable interactions in folded state	−0.837
Hydrophobic interactions	+1.105
van der Waals interactions	+0.950
Required contribution of hydrogen bonds	+0.205–+3.008
Observed net effect	+0.042

Note: N for native state and D for denatured state, taken from Creighton (1983).

$\Delta g_{i,s}$: Free energy changes for solvent contacts.

ΔW_{ei}: Free energy change for long range electrostatic interactions.

$\Delta G_{s\text{-}s}$: Free energy contributions from disulfide bonds.

The net free energy contributions of various factors that stabilize a hypothetical protein consisting of 100 amino acid residues (Creighton, 1983) have been estimated (Table 5.13).

In this evaluation, a value of 13.8–83.7 kJ/mol per residue was used as the contribution of conformational entropy to the free energy. An unfavorable interaction of 8.37 kJ/mol per residue was assumed for the folded structure. The hydrophobic free energy is calculated assuming 100 J/mol/Å^2. The van der Waal interaction due to close packing is estimated from the average enthalpy of the fusion of small, nonpolar model compounds. If the net stabilization free energy is assumed to be +41.8 kJ/mol, the hydrophobic and van der Waal interactions are insufficient. Therefore other factors must be considered to account for the free energy of +205 to $+3.01 \times 10^3$ kJ/mol. Analysis of X-ray crystallographic data for various proteins suggests that, on average, 74 hydrogen bonds exist in a protein of 100 residues. Experimental thermodynamic parameters of protein stability are available from Protherm at http://gibk26.bse.Kyutech.ac.jp/jouhou/Protherm/protherm.html (Kuman *et al.*, 2006).

5.11 REFERENCES

Abrahams, J.P., Leslie, A.G.W., Lutter, R. and Walker, J.E. (1994) *Nature*, **370**, 621–8.

Aebersold, R.H., Teplow, D.B., Hood, L.E. and Kent, S.B.H. (1986) *Journal of Biological. Chemistry*, **261**, 4229–38.

Andreeva, A., Howorth, D., Brenner, S.E. *et al.* (2004) *Nucleic Acid Research*, **32**, D226–9.

Arents, G. and Moudrianakis, E.N. (1991) *Proceedings of the National Academy of Sciences, USA*, **90**, 10489–93.

Bailey, J.M. and Shively, J.E. (1990) *Biochemistry*, **29**, 3145–56.

Barrett, G.C. (1985) *Chemistry and Biochemistry of the Amino Acids*, Chapman & Hall, London.

Bartlet-Jones, M., Jeffery, W.A., Hansen, H.F. and Pappin, D.J. (1994) *Rapid Communications in Mass Spectrometry*, **8**, 737.

Betts, L., Xiang, S., Short, S., Wolfenden, R. and Carter, C.W. Jr. (1994) *Journal of Molecular Biology*, **235**, 635–56.

Biemann, K. (1990) *Methods in Enzymology*, **193**, 886–7.

Brändén, C.-I. (1980) *Quart. Rev. Biophys*, **13**, 317–88.

Brenner, S.E., Chothia, C., Hubbard, T.J.P. and Murzin, A.G. (1996) *Methods in Enzymology*, **266**, 635–43.

Brown, R.S. and Lennon, J.J. (1995) *Analytical Chemistry*, **67**, 1998.

Buehner, M., Ford, G.C., Moras, D. *et al.* (1973) *Proceedings of the National Academy of Sciences, USA*, **70**, 305–24.

Carter, C.W. Jr. (1993) *Annual Reviews in Biochemistry*, **62**, 715–48.

Casper, D.L.D. and Klug, A. (1962) *Cold Spring Harbor Symposium in Quant. Biology*, **27**, 1–24.

Chait, B.T., Wang, R., Beavis, R.C. and Kent, S.B.H. (1993) *Science*, **262**, 89.

Chan, A.W., Hutchinson, E.G., Hanis, D. and Thanton, J.M. (1993) *Protein Science*, **2**, 1574–90.

Chou, K.-C. (1995) *Proteins: Structure, Function, and Genetics*, **21**, 319–44.

Colman, P.M., Varghese, J.N. and Laver, W.G. (1983) *Nature*, **303**, 41–4.

Creighton, T.E. (1983) *Biopolymer*, **22**, 49.

Creighton, T.E. (1993) *Proteins: Structures and Molecular Properties*. 2nd edn, W.H. Freeman, New York.

Davis, M.T. and Lee, T.D. (1992) *Protein Science*, **1**, 935–44.

Dill, K. and Chan, H.S. (1997) *Nature Stuit. Biol.* **4**, 10–19.

Edman, P. and Begg, G. (1967) *European Journal of Biochemistry*, **1**, 80–91.

Efimov, A. (1995) *Journal of Molecular Biology*, **245**, 402–15.

Egelman, E.H., Francis, N. and DeRosier, D.J. (1982) *Nature*, **298**, 131–5.

Emmett, M.R. and Caprioli, R.M. (1994) *Journal of the American Society of Mass Spectrometry*, **5**, 605–13.

Eng, J., McCormick, A.L., and Yates III, J.R. (1994) *Journal of the American Society of Mass Spectrometry*, **5**, 976–89.

Faber, G.K. and Petsko, G.A. (1990) *Trends in Biochemical Sciences*, **15**, 228–34.

Fersht, A. (1999) *Structure and Mechanism in Protein Science*, W.H. Freeman, New York.

Findley, J.B.C. and Geisow, M.J. (eds) (1989*) Protein Sequencing: A Practical Approach*, IRL Press, Oxford, UK.

Harrison, S.C., Olson, A.J., Schutt, C.E. *et al.* (1978) *Nature*, **276**, 368–73.

Hancock, W.S. (1984) *Handbook of HPLC for the Separation of Amino Acids, Peptides, and Proteins*, CRC Press, Boca Raton, FL.

Hewick, R.M., Hunkapiller, M.W., Hood, L.E. and Dreyer, W.J. (1981) *Journal of Biological Chemistry*, **256**, 7900–97.

Hunkapiller, M.W. and Hood, L.E. (1978) *Biochemistry*, **17**, 2124–33.

Hunkapiller, M.W. and Hood, L.E. (1980) *Science*, **207**, 523–5.

Hutchinson, E.C. and Thornton, J.M. (1994) *Protein Science*, **3**, 2207–16.

Jones, A.T. and Liljas, L. (1984) *Journal of Molecular Biology*, **177**, 73–92.

Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992) *Nature*, **358**, 86–89.

Jurnak, F., Yoder, M.D., Pickersgill, R. and Jenkins, J. (1994) *Current Opinions in Structural Biology*, **4**, 802–6.

Karpen, M.E., de Haseth, P.L. and Neet, K.E. (1992) *Protein Science*, **1**, 1333–42.

Kinter, M. and Sherman, N.E. (2000) *Protein Sequencing and Identification Using Tandem Mass Spectrometry*, John Wiley & Sons, New York.

Kouranov, A., Xie, L., Delacruz. J., Chen, L., Westbrook, J., Bourne, P.E. and Berman, H.M. (2006) *Nucleic Acid Resear.* **34**, D302–D305.

Krämer, A. (1996) *Annual Reviews in Biochemistry*, **65**, 367–409.

Kumar, S.M.D., Bava, K.A., Gromiha, M.M., Prabakaran, P., Kitajima, K., Uedaira, H. and Sarai, K. (2006) *Nucleic Acid Resear*, **34**, D204–D206.

Leijonmark, M., Appelt, K., Badger, J. *et al.* (1988) *Proteins*, **3**, 244.

Lesk, A.M. and Chothia, C. (1984) *Journal of Molecular Biology*, **174**, 175–91.

Leszczynski, J. and Rose, G. (1986) *Science*, **234**, 849–55.

Levitt, M. and Chothia, C. (1976) *Nature*, **261**, 552–8.

Liddington, R.C., Derwenda, Z., Dodson, E. *et al.* (1992) *Journal of Molecular Biology*, **228**, 551–74.

Lupas, A.N., Ponting, C.P. and Russell, R.B. (2001) *Journal of Structural Biology*, **134**, 191–203.

Martin, A.C.R., Orengo, C.A., Hutchinson, E.C. *et al.* (1998) *Structure*, **6**, 875–84.

Matsudaira, P. (1987) *Journal of Biological Chemistry*, **262**, 10035–8.

McLafferty, F.W. (ed.) (1983) *Tandem Mass Spectrometry*, John Wiley & Sons, New York.

Michalopoulos, I., Torrance, G.M., Gilbert, D.R. and Westhead, D.R. (2004) *Nucleic Acid Research*, **32**, D251–4.

Miller, S., Janin, J., Lesk, A.M. and Chothia, C. (1987a) *Journal of Molecular Biology*, **196**, 641–56.

Miller, S., Lesk, A.M., Janin, J. and Chothia, C. (1987b) *Nature*, **328**, 834–6.

Needlman, S.B. (ed.) (1975) *Protein Sequence Determination*, Springer-Verlag, New York.

Noble, M.E.M., Verlinde, C.L.M.J., Groendijk, H. *et al.* (1991) *Journal of Medical Chemistry*, **34**, 2709–16.

Oldfield, T.J. and Hubbard, R.E. (1994) *Proteins: Structure, Function and Genetics*, **18**, 324–37.

Orengo, C., Jones, D. and Thornton, J.M. (1994) *Nature*, **372**, 631–4.

Orengo, C.A. (1999) *Protein Science*, **7**, 233–42.

Orengo, C.A. and Thornton, J.M. (2005) *Annual Reviews in Biochemistry*, **74**, 867–900.

Pai, E.F., Krengel, U., Petsko, G.A. *et al.* (1990) *EMBO Journal*, **9**, 2351–9.

Pearl, F., Todd, A., Sillitoe, I. *et al.*, (2005) *Nucleic Acid Research*, **33**, D247–51.

Ponting, C.P. and Russell, R.R. (2002) *Annual Reviews in Biophysical Biomolecular Structure*, **31**, 45–71.

Radzicka, A. and Wolfenden, R. (1994) *Science*, **265**, 936–7.

Ramachandran, G.N., Ramakrishnan, C. and Sasisekharan, V. (1963) *Journal of Molecular Biology*, **7**, 95–9.

Ramakrishnan, V. and White, S.W. (1998) *Trends in Biochemical Science*, **23**, 208–12.

Richards, F.M. and Lim, W.A. (1993) *Quarterly Reviews in Biophysics*, **26**, 423–98.

Richardson, J.S. (1973) *Proceedings of the National Academy of Sciences, USA*, **73**, 2619–23.

Richardson, J.S. (1977) *Nature*, **268**, 495–500.

Richardson, J.S., Getzoff, E.D. and Richardson, D.C. (1978) *Proceedings of the National Academy of Sciences, USA*, **75**, 2574–8.

Rossmann, M.G., Moras, D. and Olsen, K.W. (1974) *Nature*, **250**, 194–9.

Salem, G.M., Hutchinson, E.G., Orengo, C.A. and Thornton, J.M. (1999) *Journal of Molecular Biology*, **287**, 969–81.

Shakhnovich, B., Dokholyan, N., DeLisi, C. and Shakhnovich, E. (2003) *Journal of Molecular Biology*, **326**, 1–9.

Simpson, R.J., Moritz, R.L., Begg, G.S. (1989) *Analytical Biochemistry*, **177**, 221–36.

Skolnick, J., Kolinski, A. and Ortiz, A.R. (1997) *Journal of Molecular Biology*, **265**, 217–41.

Smith, B.J. (ed.) (2002) *Protein Sequencing Protocols*, 2nd edn, Humana Press, Totowa, NJ.

Smith, C.A. and Rayment, I. (1996) *Biophysics Journal*, **70**, 1590–602.

Srinivasan, R. and Rose, G.D. (1995) *Proteins: Structure, Function, and Genetics*, **22**, 81–99.

Sternberg, M.J.E. and Thornton, J.M. (1977) *Journal of Molecular Biology*, **110**, 269–83.

Sternberg, M.J.E. and Thornton, J.M. (1978) *Nature*, **271**, 15–20.

Stevens, R.C., Gouaux, J.E. and Lipscomb, W.N. (1990) *Biochemistry*, **29**, 7691–9.

Thiede, B., Salnikow, J. and Wittmann-Liebold, B. (1997) *European Journal of Biochemistry*, **244**, 750–4.

Trabi, M. and Craik, D.J. (2002) *Trends in Biochemical Science*, **27**, 132–8.

Walker, J.E., Saraste, E.M., Runswick, M.J. and Gay, N.J. (1992) *EMBO Journal*, **1**, 945–51.

Whitehouse, C.M., Drelyer, R.N., Yamashita, M. and Fenn, J.B. (1985) *Analytical Chemistry*, **57**, 675–9.

Whitford, D. (2005) *Proteins: Structure and Function*, John Wiley & Sons, Hoboken, NJ.

Widom, J. and Klug, A. (1985) *Cell*, **43**, 207–13.

Wilm, M., Shevchenko, A., Houthaeve, T. *et al.* (1996) *Nature*, **379**, 466–9.

Wilmanns, M., Hude, C.C., Davis, D.R. *et al.* (1991) *Biochemistry*, **30**, 9161–9.

Yoder, M.D., Lietzke, S.E. and Jurnak, F. (1993) S*tructure*, **1**, 241–51.

Zhang, L. and Hermans, J. (1996) *Proteins: Structure, Function, and Genetics*, **24**, 433–8.

World Wide Webs cited

CADB;	http://cluster.physics.iisc.emet.in/cadb/
CATH:	http://www.biochem.ucl.ac.uk/bsm/cath
Cn3D	http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.html
CyBase:	http://research.imb.uq.edu.au/cybase
Histone Database:	http://research.nhgri.nih.gov/histones/
KineMage:	http://kinemage.biochem.duke.edu/
PepConfDB:	http://www.peptidome.org/products/list.html
PIR-International Protein Sequence Database:	http://pir.georgetown.edu
Protein Data Bank (PDB):	http://www.rcsb.org/pdb/
Protherm:	http://gibk26.bse.Kyutech.ac.jp/jouhou/Protherm/protherm.html
TOPS:	http://tops.leeds.ac.uk/tops/
RasMol:	http://www.umass.edu/microbio/rasmol/
SCOP:	http://scop.mrc-lmb.cam.ac.uk/scop/
Swiss-Prot of Swiss Institute of Bioinformatics:	http://expasy.hcuge.ch/sprot/
UniProt consortium:	http://www.uniprot.org

CHAPTER 6

BIOMACROMOLECULAR STRUCTURE: POLYSACCHARIDES

6.1 PROPAGATION OF POLYSACCHARIDE CHAINS

6.1.1 Introduction

Polysaccharides are polymers of monosaccharides (Boons, 1998; Davis and Fairbanks, 2002). The generic term, 'glycan' for polysaccharide evolved from the generic word 'glycose' (sugar, monosaccharide). These polysaccharide names are formed by replacing the suffix –ose in the specific sugar name by the suffix—an, thus fructan for fructose polymer, xylan for polymers of xylose, galactomannans for galactose-mannose combinations. It is noted that a name such as glucan does not refer to a specific structure but signifies only that the polysaccharide is composed of glucose residues that include cellulose, starch, glycogen, laminaran or other glucose polymers. Many of the polysaccharides have common names of long standing, e.g. cellulose, starch, glycogen, chitin, pectin, inulin, haparin chondroitin and so on. Polysaccharides are recognized to have a variety of biological functions, some of which are:

- storage of the chemical energy obtained from the sun in the process of photosynthesis;
- structural material for the cell walls of plants and microorganisms and the exoskeltons of insects and other arthropods;
- protection of organisms, especially microorganisms from changes in the environment such as changes in temperature, pH and concentration of oxygen;
- adaptation and fixation of organisms to a specific environmental niche;
- protection of organisms from invasion from other organisms and viruses, and protection against unwanted destruction by the immunological process;
- alternation of biological environments to produce a desired condition such as the prevention of blood coagulation or the prevention of drying;
- the action as lubricant in the movement of muscles and joints;
- structure of skin, cartilage and corona;
- biological recognition involved in infection and immunity, cell–cell interaction;
- receptor binding and response in signal transduction.

Structural polysaccharides are almost always linear molecules, while polysaccharides that serve primarily as energy sources are commonly branched, or in some cases (e.g. starch) a mixture of linear and branched polysaccharides with the branched type predominating. In general, branched polysaccharides are easily soluble in water and have

Biomacromolecules, by C. Stan Tsai
Copyright © 2007 John Wiley & Sons, Inc.

thickening powers. However, linear molecules are excellent structural materials because they pack closely and form many intermolecular secondary interactions, which makes the structure strong, rigid and insoluble or not easily soluble in an aqueous solution.

Polysaccharides can be divided into two classes: homopolysaccharides (homoglycans) consisting of only one kind of monosaccharide and hetereopolysaccharides (heteroglycans) consisting two or more kinds of monosaccharide units. In the heteroglycans, the arrangement or sequence of the glycose units is usually in a definite, repeating pattern, rather than random. Many naturally occurring heteroglycans are AB glycans composed of repeating sequences of disaccharides. Furthermore, the linkages can be homolinkages with either an α- or β-configuration to a single position (exclusive of any branch linkages), whereas a mixture of α- and β-configurations and/or a mixture of positions results in the heterolinkages. Thus polysaccharides can have different sequences of glycose units, different sequences of glycosidic linkages and different kinds of branching. This provides a high degree of diversity for polysaccharides and their structure-function relationships (Dumitriu, 2005).

The number of monosaccharide units in a polysaccharide molecule is termed the degree of polymerization (DP). The conformation of the individual monosaccharide residues in a polysaccharide is relatively fixed. The different kinds of primary structures that result in secondary and tertiary structures give different kinds of properties such as water solubility, aggregation and crystallization, viscosity, gelation, digestibility and biological recognition.

6.1.2 Representation of glycan structures

When representing the structure of oligo- and polysaccharides, the important information is the nature of the monosaccharide residues and how they are joined together. The method of representing the structure needs to show the sequence of the monosaccharide residues, the types of glycosidic bonds, and the type of branching (Robyt, 1986). These facets of oligo- and polysaccharide structures can be simply illustrated by representing the pyranose ring as a circle with the five hydroxyl groups represented as lines extending from the circle. Those hydroxyl groups that are above the plane of the pyranose ring in the Haworth or conformational formula are represented by placing the line inside the circle, and those hydroxyl groups that are below the plane of the ring are represented by placing the line outside the circle. The orientation of the circle is the same as the Haworth or conformational formulas, with the hemiacetal hydroxyl to the right and the other hydroxyl groups at positions 2, 3, 4 and 6 following clockwise around the circle. When carbohydrate residues are combined together, only the glycosidic bond need be shown. When more than one type of monosaccharide residue are linked together to form heterosaccharide structures, they may most easily be distinguished by using a single upper case letter or possibly two letters to represent each different monosaccharide residue. For some monosaccharide residues that have distinguishing substituents, such as acetamido, carboxyl, halogen and so on, the groups can simply be added to the particular line where they are substituted on to the pyranose ring. This simplified method for representing polysaccharide structures is useful for illustrating various branching patterns. However, this representation is not practical for glycomic analysis, which prefers linear codes.

6.1.3 Toward linear code for glycans

6.1.3.1 IUPAC Nomenclature. The summary of IUPAC carbohydrate nomenclature (McNaught, 1997) is available at http://www.chem.qmw.ac.uk/iupac/. Some highlights for glycan nomenclature (Table 2.4) are given below:

1. Each symbol for a monosaccharide unit is preceded by the anomeric descriptor (a/α for α- and b/β for β-anomer) and the configuration symbol (D- or L-). The ring size is indicted by an italic *f* for furanose or *p* for pyranose. The locant (location) of the linkage is given in parentheses between the symbols with an arrow indicating the glycosidic linkage in the direction from the anomeric carbon atom toward the hydroxyl group. A double-headed arrow indicates a linkage between the anomeric positions.
2. Oligosaccharide without a free hemiacetal group (without reducing end) is named as glycosyl glycoside. For example, α-D-galactopyranosyl-(1→6)-α-D-glucopyranosyl-β-D-fructofuranoside, or α D-Gal*p*-(1→6)-α-D-Glc*p*-(1↔2)-β-D-Fru*f* represents raffinose.
3. Oligosaccharides with a free hemiacetal group (with reducing end) are depicted with the reducing glycose residue on the right and the nonreducing glycosyl group on the left. Internal sugar units are glycosyl residues. For example, β-D-glucopyranosyl-(1→4)-β-D-glucopyranosyl-(1→4)-D-glucopyranose, β-D-Glc*p*-(1→4)-β-D-Glc*p*-(1→4)-D-Glc*p* represents cellotriose.
4. Branched oligosaccharides are named by enclosing the branches in square brackets. In a branched chain, the longest chain is regarded as the parent. If two chains are of equal length the one with lower locants at the branch point is preferred. For example, 5-*N*-acetyl-α-neuraminyl-(2→3)-β-D-galactopyranosyl-(1→3)-α-L-fucopyranosyl-(1→4)-2-acetamido-2-deoxy-D-glucopyranose, and α-Neu*p*5Ac-(2→3)-β-D-Gal*p*-(1→3)-[α-L-Fuc*p*-(1→4)]-D-Glc*p*NAc represents sialyl-Le^a trisaccharide.
5. Extended form versus short form: In the extended form, the branch connection is explicitly depicted and a short notation is derived by omitting:
   a) locants of anomeric carbon atoms;
   b) the parentheses around the locants of the linkage; and
   c) hyphens.

   For example, the extend and short forms of sialyl-Le^a trisaccharide are

   α-Neu*p*5Ac-2→3-β-D-Gal*p*-1→3-D-Glc*p*NAc
                                                  4   Neu5Acα3Galβ3[Fucα4]GlcNAc
                                                  ↑
                                       α-L-Fuc*p*-1
   Extended form   Short form

6. Homopolysaccharide (homoglycan): a general term for a polysaccharide composed of a single type of monosaccharide residue is obtained by replacing the ending -ose by -an, such as (1→4)-α-D-glucan.
7. Heteropolysaccharide (heteroglycan): For a polysaccharide containing two or more kinds of glycoses, a principle chain composed of only one type of glycose residue is cited last (as a glycan term) and the other types of residue cited as *glyco*- prefixes in an alphabetical order. If no single type of glycose residue constitutes the principle chain, all glycose residues should be cited alphabetically as *glyco*- prefixes, and the name terminates with the suffix -*glycan*, e.g. D-galacto-D-mannan.

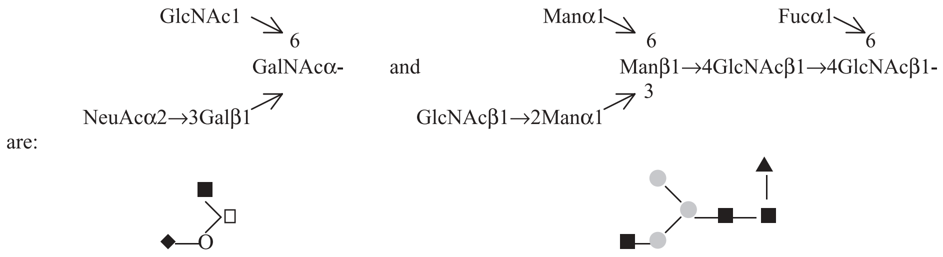
6.1.3.2 Symbol nomenclature. For symbolic representation of glycan linkage for the annotation of mass spectra, the Consortium for Functional Glycomics (http://www.functionalglycomics.org/static/consortium/) agrees that:

- Each sugar type (e.g. sugars of the same mass) should have the same shape.
- Isomers of each sugar type (e.g. Gal/Glu/Man) should be differentiated by color or by white/black/shading.
- The same color or shading should be used for derivatives of hexoses.
- Using the same shape but different orientation to represent different sugars should be avoided so that structures can be represented either horizontally or vertically.

Accordingly, symbol nomenclature (black and white version) for:
Hexoses (circles) and *N*-acetylhexosamine (square):

Galactose (white):	O	□
Glucose (black):	●	■
Mannose (gray):	●	■
Fucose (triangle)	▲	
Xylose (star)	☆	
Sialic acid (diamond)	◆	

For example, symbols for



6.1.3.3 LinearCode. The linear codes for presenting sequences of nucleic acids and proteins have been instrumental in the development of bioinformatics tools, which serve as the foundation of genomics and proteomics. In order to develop parallel tools for glycomics, a similar linear code that can be applied to represent the sequence of glycans, especially complex carbohydrates is needed. Linear Notation for Unique description for Carbohydrate Sequences (LINUCS), (Bohne-Lang *et al.*, 2001) converts structures (sequences) of complex carbohydrates to descriptive linear notations and the tool is available at http://www.dkfz.de/spec/linucs/ or http://glycosciences.de/tools/linucs/. An alternative one/two-letter symbol for the linear code (LinearCode) to represent complex carbohydrates has been proposed (Banin *et al.*, 2002). To facilitate the development and implementation of linear code(s) applicable in glycomics, the LinearCode that has been adapted by the Consortium for Functional Glycomics and offers a scheme of representing monosaccharide units and their connections in glycans will be described here.

A. Representation of monosaccharide (saccharide) units
   1. The most common structures of monosaccharides (CSM) are represented by a single letter code, as listed in Table 6.1.
   2. For the monosaccharides that are different from the common structures, they are expressed by:
      a) Stereoisomer (D or L) of CSM is indicated with an apostrophe (′), e.g. L-Glc*p* as G′.

TABLE 6.1 Linear codes of common monosaccharide units. The table is organized in the order of branch hierarchy that has been empirically determined according to the frequency in which these monosaccharides occur at the branch node

Trivial name	Common monosaccharide (CMS)	Linear code
D-Glc*p*	D-Glucose	G
D-Gal*p*	D-Galactose	A
D-Glc*p*NAc	*N*-Acetylglucosamine	GN
D-Gal*p*NAc	*N*-Acetylgalactosamine	AN
D-Man*p*	D-Mannose	M
D-Neu*p*5Ac	*N*-Acetylneuraminic acid	NN
D-Neu*p*	*N*euraminic acid	N
D-Gal*p*A	D-Galacturonic acid	L
D-Ido*p*A	D-Ioduronic acid	I
L-Rha*p*	L-Rhamnose	H
L-Fuc*p*	L-Fucose	F
D-Xyl*p*	D-Xylose	X
D-Rib*p*	D-Ribose	B
L-Ara*f*	L-Arabinofuranose	R
D-Glc*p*A	D-Glucuronic acid	U
D-All*p*	D-Allose	O
D-Api*p*	D-Apiose	P
D-Fru*f*	D-Fructofuranose	E

Notes: 1. All the monosaccharides are in their pyranose form unless otherwise noted.

2. Three common uronic acids and *N*-acetyl glycoses are considered as CMS.

3. The monosaccharides are organized in the order of branch hierarchy that is used to assign the branch chain, i.e. the chain begins with the lower monosaccharide in the hierarchy is assigned to be the branch chain.

b) Alternative ring structure (furanose or pyranose) of CSM is indicated by a caret (∧), e.g. D-Glc*f* as G∧.

c) A monosaccharide that differs in both stereoisomerism and ring structure from CSM is indicated by a tilde (~), e.g. L-Glc*f* as G~.

3. Modifications are represented by adding square brackets that include the connecting position of the modification to the monosaccharide code, followed by the modification symbols (Table 6.2), e.g. D-Glc*p*-6-phosphate is written as G[6P]. Exceptions include monosaccharides with common modifications such as N-acetyl-D-glucosamine (GN), *N*-acetyl-D-galactosamine (AN) and *N*-acetylneuraminic acid (NN).

B. Connection of monosaccharide units

1. Lower case symbols are used to represent connections (glycosidic linkages). Two components (anomer and linkage position) that describe a connection between two monosaccharide units are indicated. The letters a- and b-, representing α- and β-anomers, are followed by the number corresponding to their connection position.
2. The linear code is read from right to left consistent with the convention in which carbohydrates are read from the right, i.e. nonreducing (NR) end to left, i.e. reducing (R) end.
3. Linear glycans: A linear glycan refers to an unbranched and noncyclic string of monosaccharide units (see below for heteroglycans with repeating units) e.g.

TABLE 6.2 Linear codes of common modifications in glycans

Modification	Linear code
N-Acetyl	N
O-Acetyl	T
Deacetylated N-acetyl	Q
Ethanolaminephosphate	PE
Inositol	IN
Methyl	ME
Phosphate	P
Phosphocholine	PC
Pyruvate	PYR
Sulfate	S
Sulfide	SH
2-Aminoethylphosphonic acid	EP

α-D-Glc*p*-(1→4)β-D-Gal*p*-(1→3)-α-D-Glc*p*-(1→4)β-D-Gal*p*-(1→3)α-D-Neuraminic acid is written as: Ga4Ab3Ga4Ab3Na.

4. Branched glycans: The branches are written within the parenthesis ( ). The designation of branch chain versus backbone chain follows:
   a. When the chains begin with identical monosaccharides, the chain connected to the higher position is considered the branch, e.g. in IUPAC-IUBMB

αD-NeuNAc-2→3-βD-Gal-1→4-βD-glcNAc-1
↓
4
αD-Man-1→3-βD-Man-1→4-αD-GlcNAc
2
↑
βD-GlcNAc-1

is written as: GNb2(Nna3AB4GNb4)Ma3Mb4Gna in the LinearCode.

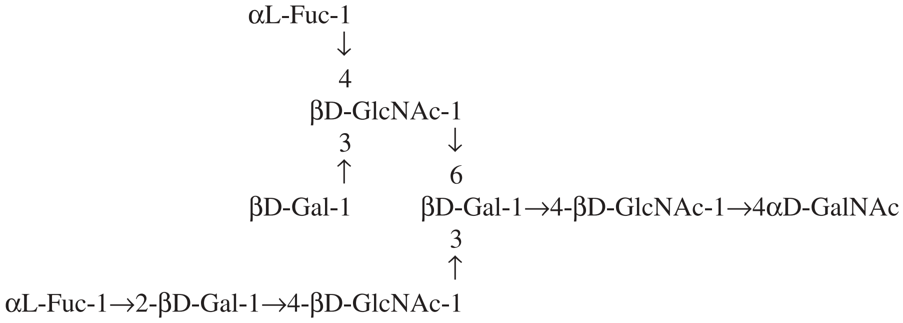
   b. When the chains begin with different monosaccharides, the chain beginning with the monosaccharide lower in the branch hierarchy is considered the branch:

βD-Gal-1→4-βD-glcNAc-1
↓
4
βD-GalNAc-1→4-βD-Glc
3
↑
αD-NeuNAc-2

is written as: Ab4Anb4(Nna3)Anb4Gb.

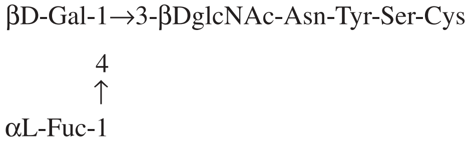
   c. Modifications do not change the hierarchy, except those considered as CSM.
   d. For multiple branches, the branch/backbone decision is determined first by the branch hierarchy, then by the connection positions:

is written as: Fa2Ab4GNb3(Ab3(Fa4)GNb6)Ab4GNb4Ana.



$$\begin{array}{lll}
 & \alpha\text{L-Fuc-1} & \\
 & \downarrow & \\
 & 4 & \\
 & \beta\text{D-GlcNAc-1} & \\
 & 3 & \downarrow \\
 & \uparrow & 6 \\
 & \beta\text{D-Gal-1} & \beta\text{D-Gal-1}\rightarrow 4\text{-}\beta\text{D-GlcNAc-1}\rightarrow 4\alpha\text{D-GalNAc} \\
 & & 3 \\
 & & \uparrow \\
\multicolumn{3}{l}{\alpha\text{L-Fuc-1}\rightarrow 2\text{-}\beta\text{D-Gal-1}\rightarrow 4\text{-}\beta\text{D-GlcNAc-1}}
\end{array}$$

5. Heteroglycans containing repeating units: The repeating units are expressed inside parenthesis, '{n}', where n represents the number of repeats. For example, amylose, which is the glucan of D-Glc joined by α-1,4 linkage is written as {nGa4}. If the repeating units are not connected head to tail, the monosaccharide at which the unit is connected is marked between two dashes '- -'.
6. Glycoconjugates/glycosides: Glycans connected via its reducing end in:
   a. Glycoproteins: Amino acid sequences (in single letter code) are written after a semicolon, ';' e.g.:



$$\begin{array}{ll}
\multicolumn{2}{l}{\beta\text{D-Gal-1}\rightarrow 3\text{-}\beta\text{DglcNAc-Asn-Tyr-Ser-Cys}} \\
 & 4 \\
 & \uparrow \\
\multicolumn{2}{l}{\alpha\text{L-Fuc-1}}
\end{array}$$

   is written as: Ab3(Fa4)GNb;NYSC.

   b. Lipopolysaccharides: Lipid moieties (e.g. C, D, IPC and DAG for ceramide, sphingosine, inositolphosphoceraminde and diacylglycerol respectively) are written after a colon, ':' e.g. βD-Glc bound to sphingosine is written as Gb:D.
   c. Glycosides: The complete name of an aglycon is written after a number symbol, '#' e.g. 4-nitrophenyl αD-glucoside as Ga#4-nitrophenol.

C. Unknown/uncertain elements

The following linear code symbols are used:

Symbol	Element
?	Single unknown saccharide unit
*	Entire string of saccharide units unknown
/	Separation of two possibilities for a given glycan chain
//	Separation of two uncertain identities of saccharide units

6.2 SEQUENCE ANALYSIS OF POLYSACCHARIDES: PRIMARY STRUCTURE

For polysaccharide structures, the same definition applies for linear chains, but the manner in which the various monosaccharide residues are joined is much more complex than for the corresponding combination of nucleotides or amino acid residues, due to the greater number of substitution positions possible in a monosaccharide molecules. Thus in order to define the primary structure of a polysaccharide fully, it is essential to state the identity of all monosaccharide residues, the sequence of these residues, their position and anomeric configuration and the position of any other substituents. Thus analysis of the primary structure of polysaccharides is complicated by a number of parameters, such as:

- The nature, order and ring conformation of individual monosaccharides.
- The anormericity (α- or β-linkage) of individual glycosidic bonds.
- Substitution patterns and branching points.
- Absolute stereochemistry of individual residues (D- or L-).
- The nature and location of chemical substituents (e.g. acetyl, methyl, phosphate, and sulfate) on a given monosaccharide.

The primary structures of polysaccharides tend to be regular, involving large blocks in which a single residue or a single sequence is repeated. These repeating sequences have been determined for many polysaccharides and other carbohydrate-containing macromolecules.

No one method available for structural analysis will provide sufficient data to allow the structure of oligo/polysaccharide to be defined in terms of its component monosaccharide units, the inter-unit linkages, and the sequence in which the units are linked. This information can only be obtained using a number of different techniques in conjunction with one another, such as chemical, enzymatic and spectrometric methods.

6.2.1 Hydrolysis to constituent monosaccharides

Once a pure sample of the polysaccharide has been obtained, the first step in elucidating its structure is to identify and estimate the component monosaccharides. The conditions for hydrolysis must be controlled such that complete hydrolysis is achieved with little or no degradation of the monosaccharide units. Use of more than one set of hydrolysis conditions may be necessary. The ease of hydrolysis of different linkages and the stability of the various monosaccharides means that the optimum conditions for each polysaccharide has to be determined. Polysaccharides containing furanose and 2 deoxy-hexoses or pentose are more readily hydrolyzed than those containing hexuronic acid or 2-amino-2-deoxyhexoses, with hexose-containing polysaccharides being intermediate. A general procedure employs 10% (v/v) concentrated hydrochloric acid or trifluoroacetic acid (TFA) at 100–120°C in a sealed tube for 0.5–4 h. For hexose-containing polysaccharides, 1 M sulfuric acid at 100°C for 4 h has been found to be appropriate, whereas the use of 0.25 M sulfuric acid at 70°C is recommended for pentose-containing polysaccharides. Degradation frequently occurs in direct hydrolysis whatever conditions are used. For example, in the case of glycosaminoglycans, 4 M hydrochloric acid at 100°C for 9 h is necessary to liberate all the 2-amino-2-deoxyhexose residues, but under such conditions the majority of the hexuronic acid residues are decomposed.

The hydrolysate (monosaccharide mixture) is analyzed by chromatographic methods. The ion exchange chromatography of the borate complexes of neutral monosaccharides is the most common method, which is used as the basis of fully automatic carbohydrate analysis.

6.2.2 Chemical methods

Some chemical methods (Bouveng and Lindberg, 1960), which have been essential to structural analyses of carbohydrates for many years will be briefly described. Although widely used before, these methods have largely been replaced by enzymatic and spectrometric methods in recent years, because they suffer from the disadvantages of time consuming, incomplete reliability, and low sensitivity.

6.2.2.1 *Fractional analysis.* The sequence of glycoses and their linkages in a glycan can be determined by analyzing partially hydrolyzed fragments known as fractional analysis. Because of the differences in the stability of the glycosidic bonds of the various monosaccharides, it is often feasible to obtain information with regard to the glycose sequence and linkages by characterizing mono- and oligosaccharides released under mild (controlled) acid hydrolysis (e.g. 0.05–0.10 N H_2SO_4 at 100°C). For hexopyranoses, the 1→6 linkages are more stable than the secondary linkages of 1→2, 1→3, or 1→4, and β-linkages are more stable than the corresponding α-linkages. Partially hydrolyzed samples are taken at time intervals for analysis (McGinnis and Fang, 1980). Generally, HPLC and LC/MS may provide structural identity of the isolated hydrolysis fragments.

6.2.2.2 *Methylation analysis.* Methylation analysis was developed to provide information about the positions of attachment of the glycosidic linkages of monosaccharide residues in glycans. The sample is dissolved in Hakomori reagent (dry dimethylsulfoxide with sodium hydride) and then methylated with methyl iodide (Hakomori, 1964). The methylated sample is acid hydrolyzed, and the methylated monosaccharides are analyzed. For the GC/MS analysis, the methylated monosaccharides are reduced and acetylated. The position of glycosidic linkages in the glycose is deduced by the types of methylated monosaccharides that are produced. For example, a glucan linked by 1→4 with 1→6 linkages produces three types of methylated glucoses: 2,3,6-tri-*O*-methyl-D-glucose from the inner residues of the main chains; 2,3-di-*O*-D-glucose from the branched residue and 2,3,4,6-tetra-*O*-methyl-D-glucose from the end residues of the chains.

6.2.2.3 *Periodate oxidation.* Periodate oxidation has been used as a collaborative approach to determine the position of the glycosidic linkages in glycoses. Periodate stoichiometrically cleaves C—C bonds of vicinal hydroxyl carbons (glycol cleavage). The reaction is carried out with periodic acid and its salts in an aqueous solution (pH 3–5). One mole of periodate is consumed per glycol (diol) structure. Oxidation of a primary hydroxyl group adjacent to a secondary hydroxyl group leads to the formation of formaldehyde, whilst vicinal triol groups yield formic acid. The consumption of periodate (oxidation) is monitored spectrophotometrically at 290 nm, using $\varepsilon_{290} = 0.22$ mM periodate/cm. The periodate consumption and product formation allows the differentiation among the linkages (Table 6.3):

6.2.3 Enzymatic methods

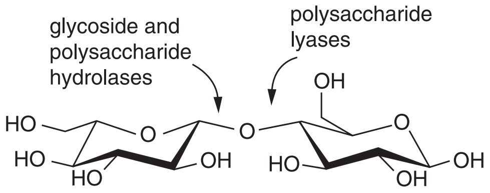
Enzymes can be used to provide both qualitative and quantitative determination of carbohydrates. Hydrolysis by enzymes provides an alternative method for the controlled hydrolysis of polysaccharides. The basis of enzymatic sequencing is to evaluate the susceptibility of glycans to a series of sequence-grade glycosidases of defined specificity. The information obtained is not limited to that obtainable by analysis of the hydrolysis fragments because the specificity of enzyme action, a specificity based on type of monosaccharide and type of linkage, leads to significant data being obtained by a process of elimination, from enzyme-resistant structures and partially hydrolyzed structures. Usually other methods, such as chromatographic analysis, MS and NMR spectroscopies are employed further to establish the structures.

TABLE 6.3 Periodate oxidation of linked hexopyranose residues

Hexose linkage		Observation per residue
1,2-linkage:	Inner residues	One periodate consumed.
	Terminal reducing residue	Three periodate consumed/one formaldehyde (C^6) and two formic acid (C^4 and C^5) formed.
1,3-linkage:	Inner residues	No reaction.
	Terminal reducing residue	Three periodate consumed/one formaldehyde (C^6) and two formic acid (C^1 and C^5) formed.
1,4-linkage:	Inner residues	One periodate consumed.
	Terminal reducing residue	Three periodate consumed/one formaldehyde (C^6) and two formic acid (C^1 and C^2) formed.
1,6-linkage:	Inner residues	Two periodate consumed and one formic acid (C^3) formed.
	Terminal reducing residue	Four periodate consumed/four formic acid (C^1, C^2, C^3 and C^4) formed.

Note: For glycans of hexopyranoses, the terminal non-reducing unit consumes 2 moles of periodate with the formation of one mole of formic acid (C^3).

The enzymatic hydrolysis of glycosidic linkages by hydrolases involves scission of the glycosyl–oxygen bond. However, a number of enzymes known as eliminases or lyases (usually of bacterial origin) react by a different mechanism and cause cleavage of the oxygen–aglycone bond in acidic polysaccharides (such as pectins), producing unsaturated hexuronic acid units.



The enzymes, which hydrolyze polysaccharides, are divided into two groups, *endo*- and *exo*-polysaccharide hydrolases (endoglycosidases and exoglycosidases). Endoglycosidases are specific for linkage and monosaccharide residue and cause random fragmentation of homopolysaccharides to give a homologous series of oligosaccharides, e.g. α-amylase, which gives a random series of D-Glc oligomers on reaction with amylose. Exoglycosidases are specific for monosaccharide unit and stereochemistry at C1 but do not differentiate between residues attached glycosidically at C1. They cleave polysaccharides by sequential removal of residues from one end of the molecule, usually the nonreducing end, e.g. β-amylase, which removes maltose units sequentially from amylose, producing an almost quantitative amount of maltose if the reaction goes to completion.

A number of glycosidic hydrolases have been produced in sufficiently pure form to allow the development of a method of determination of monosaccharide sequences based on these enzymes (Table 6.4). These enzymes will remove specific monosaccharide units linked by specific linkages from the nonreducing end of a polysaccharide. For example, β-galactosidase will remove D-galactosyl residues linking β-glycosidically to polysaccharide. The use of sequential enzymatic hydrolysis is a well-established technique, particularly for the analysis of the carbohydrate residues of macromolecules.

TABLE 6.4 Enzymes commonly used in glycan structure analysis

Enzyme	Source	Specificity
Exoglycosidases		
α-D-Glucosidase	*Saccharomyces cerevisiae*	Glcα1→4
β-D-Glucosidase	Almond emulsin	Glcβ1→4
α-D-Galactosidase	Green coffee bean	Galα1→3,4,6
β-D-Galactosidase	*Escherichia coli*	Galβ1→4Glc
	Streptococcus pneumoniae	Galβ1→4
	Jack bean	Galβ1→6>4>>3
α-D-Mannosidase	*Aspergillus phoenicis*	Manα1→2
	Jack bean	Manα1→2,3,6
β-D-Mannosidase	*Helix pomatia*	Manβ1→4
β-*N*-Acetyl-D-hexosaminidase	Jack bean	Glc(Gal)NAcβ1→2,3,4,6
	S. pneumoniae	Glc(Gal)NAcβ1→2,3
α-*N*-Acetyl-D-galactosaminidase	Chicken/porcine liver	GalNAcα1→
α-L-Fucosidase	Bovine epididymis	Fucα1→6>2,3,4
	Almond emulsin	Fucα1→2/3,4
β-D-Xylosidase	*Charonia lampas*	Xylβ1→2
α-D-Sialidase	*Archrobacter ureafaciens*	NeuNAcα2→6>3
	Clostridium perfringens	NeuNAcα2→3,6
	Newcastle disease virus	NeuNACα2→3
Endoglycosidases		
α-Amylase	Pig pancreas,	↓
	Bacillus amyloliquefaciens	Glcα1→4Glc
Endoglycosidase H	*Streptomyces plicatus/griseus*	↓
		(Man)$_n$-GlcNAc-GlcNAc
Endoglycosidase F1-3	*Flavobacterium meningosepticum*	
Endo-β-D-galactosidase	*Escherichia freundii*	
Endoglycosidase D	*S. pneumoniae*	
Endo-α-D-sialidase	KIF phage in *E. coli*	

6.2.4 Spectrometric methods

6.2.4.1 Polarimetric measurement. The configuration of glycosidic linkages can be estimated by the determination of the specific optical rotation of the glycose solution. If the saccharide has a relatively high positive rotation $[\alpha] \geq +100^\circ$, the glycose probably has a high number of α-linkages, and if the specific rotation is relatively low, $[\alpha] \leq 10^\circ$, it probably has a high number of β-linkages. For examples:

Equilibrium specific rotation, [α] (°)

	α-linkage	β-linkage
αD-Glucose	+113	
βD-Glucose		+19
Maltose, (αD-Glc)$_2$	+130.4	
Cellubiose, (βD-Glc)$_2$		+35
Maltotriose, (αD-Glc)$_3$	+160	
Celluotriose, (βD-Glc)$_3$		+22
Starch, (αD-Glc)$_n$	+220	
Oxidized cellulose, (βD-Glc)$_n$		−45

6.2.4.2 Mass spectrometric analysis. Mass spectrometry (Caprioli *et al.*, 1996), particularly in combination with separation/purification techniques such as capillary electrophoresis or HPLC (Millner, 1999), is the most commonly used method for the sequence analysis of glycans. Of the ionization methods for MS analysis of oligo/polysaccharides, MALD- and ESI-MS provide rapid, sensitive and essential structural information (Reinhold *et al.*, 1996), although the number of constituent isobaric monosaccharides tends to be limited (Table 6.5). Thus an ultimate identification of the monosaccharide structures (e.g. diastereoisomers, anomers and conformers among hexoses) may require additional experimentations such as enzymatic, NMR or chemical methods. For an unsubstituted glycan, the measured glycan mass is given by the sum of the monosaccharide residue mass plus the mass of the end group (OH and H, if underivatized glycans) and the mass of the adduct that is needed to ionize the molecule (usually H or Na).

The types of fragmentation observed under different ionization conditions tend to be similar. Two common types of fragmentation are glycosidic cleavages that break a bond linking two sugar residues and cross-ring cleavages that involve the breaking of two bonds. Glycosidic cleavages provide information on sequence and branching whereas the cross-ring cleavages reveal more details on linkage and bonds. In glycosidic cleavages, a unique series of fragment ions is produced from glycans and the nomenclature for describing the fragmentation and m/z ions for the glycosidic cleavages is given in Figure 6.1. Ions generated by the glycosidic cleavages that retain the charge at the reducing terminus are designated Y and Z (numbering from the reducing end), whereas ions with the charge at the non-reducing terminus are B and C (numbering from the non-reducing end). The corresponding ions generated by the cross-ring cleavages are X and A respectively.

Structural information can be obtained by MS–MS techniques. The first MS (MS-1), such as ESI, separate/isolate the degradation product and the second MS (MS-2), such as collision-induced dissociation (CID), provides the necessary product identification. The suitable precursor ions for MS–MS generally comprise B_n-type ions, which are often form good yields due to the facile cleavage of the glycosidic bond under a variety of ionization conditions.

Chemical derivatization is often needed for successful analysis and the nature of the derivative can be directed so as to induce fragmentation that may yield structural details

TABLE 6.5 Residue masses of common monosaccharides and their derivatives

Glycose	Residue Formula	Average Resid mass	Resid mass	Methyl No. Me	Resid mass	Acetyl No. Ac
Deoxy-pentose	$C_5H_8O_3$	116.117	130.144	1	158.154	1
Pentose	$C_5H_8O_4$	132.116	160.170	2	216.191	2
Deoxyhexose	$C_6H_{10}O_4$	146.143	174.197	2	230.216	2
Hexose	$C_6H_{10}O_5$	162.142	204.223	3	288.254	3
Hexosamine	$C_6H_{11}N\ O_4$	161.157	217.265	4	287.269	3
N-Ac-Hexosamine	$C_8H_{13}N\ O_5$	203.179	245.276	3	287.269	3
Hexuronic acid	$C_6H_8O_6$	176.126	218.207	3	260.200	2
N-Ac-NeuA	$C_{11}H_{17}N\ O_8$	291.258	361.392	5	417.370	3
N-Glycol-NeuA	$C_{11}H_{17}N\ O_9$	307.257	391.419	6	475.406	4

Notes: 1. Abbreviations used are: Ac, acetyl; Me, methyl; NeuA, neuraminic acid; resid, residue.

2. Average residue mass (resid mass) is given, based on C = 12.011, H = 1.00794, O = 15 9949

3. The mass of the intact glycan can be obtained by addition of the residue masses plus the figure for the relevant terminal groups (18.153 for free glycan, 46.069 for methyl derivative and 102.090 for acetyl derivative. For the mass of the molecular ion, also add the mass of the adduct, i.e. 1.0079 for hydrogen, 22.9898 for sodium and 38.0983 for potassium.

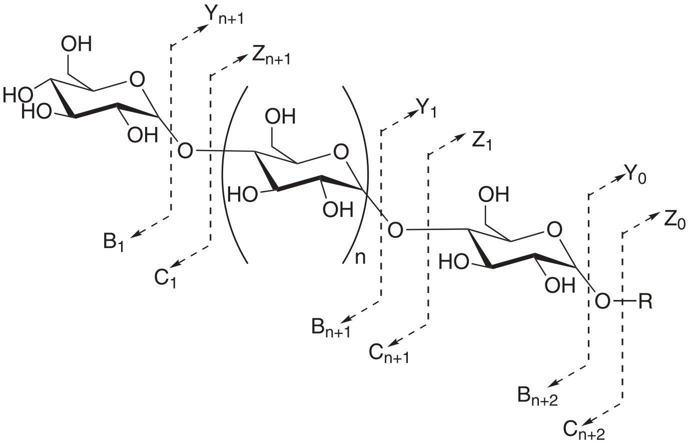


Figure 6.1 MS fragmentations of glycosidic linkages

of interest. For example, a differentiation among diastereoisomeric hexoses can be made by comparing B ions of tetra-*O*-acetyl versus tetra-*O*-(trideuterioacetyl) derivatives. The two MS B_1 ions at m/z 127 and AT m/z 169 of methyl tetra-*O*-acetyl-glucopyarnoside and -galactopyranoside are shifted and split into two components, m/z 127 → 128/129 and m/z 169 → 172/173 in tetra-*O*-(trideuterioacetyl) derivatives with m/z 172/173 ratio being about 2 : 1 for Glc- and 1 : 2 for Gal-derived B_1 ions. Table 6.6 shows three distinct types with regarding to (1) m/z 129 intensity and (2) the intensity ratio of m/z 172/173. The three distinct types for m/z 172/173 ratio are: glucose type (Glc, Ido with ratio 2 : 1), mannose type (Man, Alt with ratio 1 : 1) and galactose type (Gal, All, Gul, and Tal with ratio 1 : 2). It appears that an assignment of the three common occurring hexopyranoses, namely Glc, Gal and Man can be made by this approach.

6.2.4.3 Nuclear magnetic resonance analysis. The theory and practice of nuclear magnetic resonance (NMR) spectroscopy in the structural analysis of biomacromolecules (Evans, 1995; Roberts, 1993) will be discussed in the next Chapter (Chapter 7). In this section, the reference will be made to the three measurable parameters afforded by NMR in the context of providing sequence information of glycoses. These are:

1. Chemical shif*t* (δ in ppm) that measures the peak position(s) of the resonance nucleus and is characteristic of the given nucleus, dependent on its chemical environments in the molecular structure.
2. Signal intensity (peak height), which refers to the integration of the peaks, and measures the number of the resonance nuclei.
3. Multiplicity of resonance peaks.

The spin–spin coupling by the neighboring nuclei (n) may cause the splitting of the resonance peaks into n + 1 peaks known as multiplicity. The separation of the multiplicity of the resonance peaks is measured as the coupling constant (J in Hz). There are two types of protons (protons bound to C-atoms and O-atoms) for the ^{1}H-NMR or proton magnetic resonance (PMR) spectroscopy of glycoses. However, the hydroxyl protons are generally unsuited for structural investigations since they produce broad signals that are variable in their chemical shift and are freely exchangeable in an aqueous solutions. The protons on C_2 to C_6 (H_2—H_6) are more strongly shielded than the proton on the anomeric C-atom (H_1), which appears in a lower field. In practice, it is not usually necessary to

TABLE 6.6 Relative intensity of CID ions of methyl-tetra-O-(trideuterioacetyl) glycopyranosides

	m/z				
Ion	109	129	172	173	343
Glc type (m/z ratio, 172 : 173 ≈ 2 : 1)					
Ido*p*	100	27	31	15	2
Glc*p*	100	28	26	15	2
Man type (m/z ratio, 172 : 173 ≈ 1 : 1)					
Alt*p*	100	42	25	22	4
Man*p*	100	45	25	24	5
Gal type (m/z ratio, 172 : 173 ≈ 1 : 2)					
Tal*p*	100	77	22	38	7
Gul*p*	100	78	21	39	7
Gal*p*	100	91	20	47	8
All*p*	100	93	19	48	10

Notes: 1. Relevant peaks from CID spectra of methyl tetra-*O*-(trideuterioacetyl)-hexopyranosides of eight diastereomeric hexopyranoses are shown.

2. Extracted from Richter *et al*. (1990).

TABLE 6.7 ^{13}CMR data (ppm) for common monosaccharides and derivatives

Glycose	C-1	C-2	C-3	C-4	C-5	C-6	Other
α-Glc*p*	92.9	72.5	73.8	70.6	72.3	61.6	
β-Glc*p*	96.7	75.1	76.7	70.6	76.8	61.7	
α-Gal*p*	93.2	69.4	70.2	70.3	71.4	62.2	
β-Gal*p*	97.3	72.9	73.8	69.7	76.0	62.0	
α-Man*p*	95.0	71.7	71.3	68.0	73.4	62.1	
β-Man*p*	94.6	72.3	74.1	67.8	77.2	62.1	
α-Fru*f*	63.8	105.5	82.9	77.0	82.2	61.9	
β-Fru*f*	63.6	102.6	76.4	75.4	81.6	63.2	
α-Ara*f*	101.9	82.3	76.5	83.8	62.0		
β-Ara*f*	96.0	77.1	75.1	82.2	62.0		
α-Rib*f*	97.1	71.7	70.8	83.3	62.1		
β-Rib*f*	101.7	76.0	71.2	83.3	63.3		
α-Glc*p*NAc	92.1	55.3	72.0	71.4	72.8	61.9	23.3 (NAc)
β-Glc*p*NAc	96.2	58.0	75.2	71.2	77.2	62.0	23.5 (NAc)
α-Gal*p*NAc	92.2	51.4	68.6	69.7	71.6	62.4	23.2 (NAc)
β-Gal*p*NAc	96.5	54.9	72.3	69.0	76.3	62.2	23.4 (NAc)
α-Neu*p*5Ac	174.1	101.6	41.0	69.0	52.9	73.4	69.6, 72.6, 63.6
β-Neu*p*5Ac	176.1	101.4	40.8	67.1	53.1	71.1	69.5, 71.1, 64.5

Notes: 1. Taken from Bock *et al*. (1983)

2. Others are NAc-methyl of Glc*p*NAc and Gal*p*NAc and C-7 to C-9 for Neu*p*5Ac.

analyze the entire spectrum due to the complexity of PMR of oligo/polysaccharides. The most important diagnostic signals are in the anomeric region 4.5–6.5 ppm.

However, the proton-decoupled ^{13}C-NMR (CMR) spectrum of carbohydrates, which gives a signal for each of the specific types of carbons present, may furnish a wealth of information on the chemical structures of oligo/polysaccharides. The CMR spectra show distinctive signals for the various carbons that are characteristic for the structures of glycans (Table 6.7). In general, the C-1 signals for the α-configured glycosidic linkages range between 99 and 102 ppm, while the C-1 signals for the β-configured glycosidic linkages are more down field at 102–105 ppm (Gorrin, 1981). The NMR spectroscopy provides the best method for determination of the anomeric configuration of oligo/polysaccharides. Substitution for the hydroxyl groups attached to the carbohydrate carbons usually produces a shift in resonance signals.

6.3 CONFORMATION: SECONDARY AND TERTIARY STRUCTURES OF POLYSACCHARIDE CHAINS

There are two possible types of strainless six-member pyranose rings, the boat (B) and the chair (C) forms and of these, the chair is usually preferred as there are fewer interactions across the ring between substituents. The rings can be defined as $^4C_1$ if C4 is above the plane (described by C2, C3, C5 and O5) and C1 is below it, or as $^1C_4$ for the alternative ring. β-D-Glucopyranose has all substituents (OH) that are equatorial in the $^4C_1$ conformation, a possible reason why D-Glc is a common monosaccharide in nature.

For the furanose ring, which is adopted by some hexoses and pentoses, the five-member ring is only slightly puckered and exists in three forms, the more common envelope (E) and the less common twist (T). For the more stable envelope forms, ^3E and E_3, the conformation is defined by the C3 atom that is above or below, the plane (described by C1,C2, C4 and O4).

Despite the variety of different monomer units and type of linkages present in polysaccharide chains, the conformational possibilities are limited. The geometry of the individual sugar rings in a polysaccharide is essentially rigid but the relative orientations of component residues about the glycosidic linkage determine the overall conformation of the polysaccharide. Two torsion (rotation) angles are required in order to define the glycosidic bond between two monosaccharide residues. The angle ϕ is about the bond from the anomeric carbon atom to the oxygen atom that joins the two residues, the angle ψ is about the bond from the glycosylated oxygen atom of first residue to the carbon atom of the second residue, and ω is about the exocyclic carbon–carbon bond. The range of values obtained for the rotational angles ϕ, ψ and ω is severely restricted by steric hindrance between the adjacent residues. The restrictions are greatest for glycosidic linkages involving axial groups, and for residues containing bulky substituents in equatorial positions adjacent to the glycosidic linkage. Possible values for ϕ and ψ (taken as 0° when the two midplanes of the sugar rings are coplanar) for cellulose (β-1,4-D-glucan) with probable steric considerations show that these angles are constrained to an extremely narrow range placing the monomer units in an almost completely extended conformation (Reed and Kerrett, 1968). Each D-Glc unit is flipped over 180° from the previous one with the plane of the rings propagating in a slight zig-zag fashion in a strand sometimes referred to as a ribbon. All hydroxyl groups lie in equatorial positions to form hydrogen bonds with neighboring chains in cellulose fibers. One significant difference between the two β-D-glucans, cellulose and chitin, is the arrangement of the chains in either parallel or antiparallel direction. Natural cellulose seems to occur only in parallel arrangements. However, chitin can occur in three forms, i.e. all parallel, all antiparallel and mixed parallel and antiparallel arrangements.

However, the residues joined together by the glycosidic linkage can rotate around the bonds of the linkage to give different chain conformations. The most commonly observed secondary and tertiary structures of polysaccharides are the single and double helical structures. In starch and glycogen (α-D-glucans), the main $\alpha(1 \rightarrow 4)$ chains tend to undergo helical coiling. Each chain contains six D-Glc units per turn and the two chains may be arranged in either parallel or antiparallel directions as defined from NR (non-reducing) end $\rightarrow$ R (reducing) end of the chain.

The structure difference between amylose and cellulose, which completely alters the properties of the polymeric glucans is that the D-glucopyranose units are linked by $\alpha(1 \rightarrow 4)$ glycosidic bonds in amylose whereas the linkages are $\beta(1 \rightarrow 4)$ in cellulose. The conformational difference between these two structures; is that the $\alpha(1 \rightarrow 4)$ linkages of amylose are naturally bent, conferring a gradual turn to the polymeric chain, which results in the helical conformation (Figure 6.2A). The most stable conformation about the $\beta(1 \rightarrow 4)$ linkage involves alternating 180° flips of the glucose units along the chain so that the chain adopts a fully extended conformation referred to as an extended ribbon (Figure 6.2B). Juxtaposition of several such chains permits efficient interchain hydrogen bonding, the basis of much of the strength of cellulose. The flattened sheets of the chains lie side by side and are joined by hydrogen bonds. These sheets are laid on top of one another in a way that staggers the chains.

The repeating sequences in primary structures of polysaccharides lead to regular patterns in secondary structures, which lead to sterically regular gross conformations aided by favorable non-covalent interactions between hydroxyl and other functional groups if present (e.g. amino, carboxyl, sulfate, and phosphate groups). Irregularities in primary and secondary structures and large branched structures inhibit tertiary structure formation, whilst such external perturbations as changes in temperature and ionic concentrations can cause changes in the tertiary structure. A useful concept that has been used to describe the overall chain conformation is to regard any conformation as a helix and specify two parameters, namely the number (n) of monomer residues per helix turn and the projected length

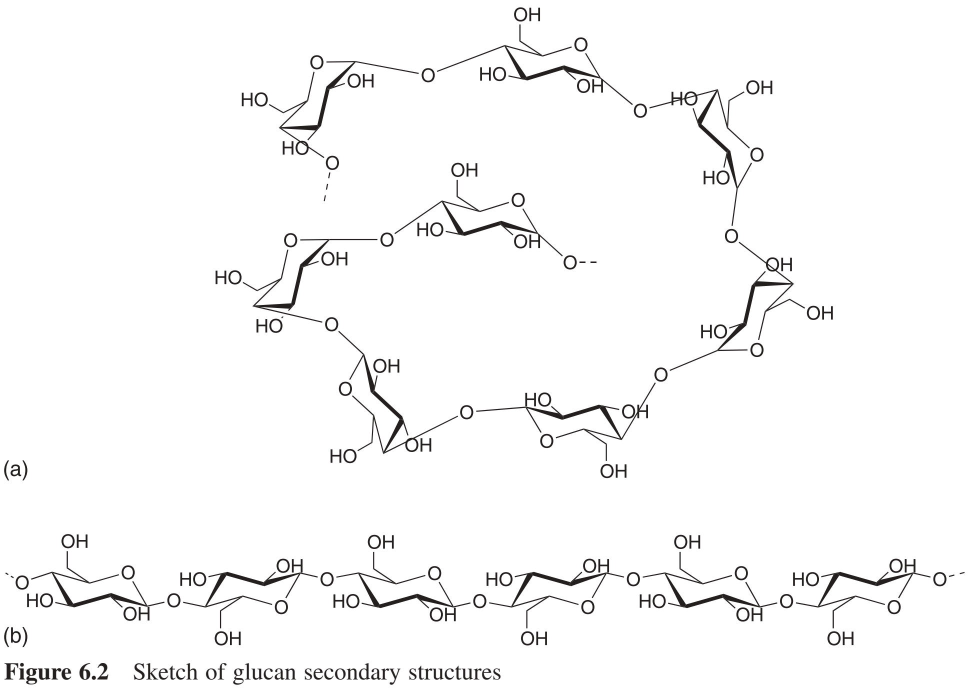


Figure 6.2 Sketch of glucan secondary structures
Secondary structures of D-glucopyranose units connected by the α-1,4-linkage (A) and β-1,4-linkage (B) respectively are schematically represented (not scaled to the actually dimension) to show the directionality (helical *versus* linear) of the structures.

(h) of each monomer residue on the helix axis. The allowed conformations of homopolysaccharides have values of n and h that fall into ranges that allow four distinct types to be identified. Type A (e.g. 1 → 4-β-D-glucan, 1 → 3/4-α-D-galactans, 1 → 4-β-D-mannan) is the extended ribbon structure with a value for n of 2 to ± 4 (negative values indicate left-handed structures) and h is close to the absolute length of the residue. Where values of n cover a wider range (n = 2 to ±10) and h approaches zero, the type B conformation (a normal helix) is obtained, such as 1 → 4-α-D-glucan, 1 → 3/4-β-D-galactans, 1 → 2/4-α-D-mannan, 1 → 4-α-D-xylan. Type C is a crumpled ribbon conformation, and examples such as 1 → 2-α/β-D-glucans, 1 → 2-α/β-D-galactans, 1 → 2-α/β-xylans are characterized by many clusters between nonadjacent sugar residues in the primary structure. The fourth type, type D is a flexible coil including all 1,6-disubstitued homoglycans. These homoglycans have more flexibility due to the extra bond that separates the rings in (1 → 6)-linked polysaccharides.

Tertiary structure in polysaccharides involves folding of the helical secondary structures. Quaternary structure is formed by association of individual helical or folded polysaccharides. This aspect of structure is frequently referred to as the subunit phenomenon and involves the aggregation of a number of chains by noncovalent bonds. The aggregation of polysaccharide chains can be between like glycan molecules, such as the interaction between cellulose chains to give the structural features of plant cell walls, or between unlike glycan molecules, such as the interaction between xanthan helices with the unsubstituted regions of the backbone of glucomannans or galactomannans. Carbohydrate Structure Suite (CSS) at http://www.dkfz.de/spec/css/ compiles and analyzes carbohydrate 3D structures derived from the PDB. These analytical tools are also available at Glycosciences (http://www.glycosciences.de/tools/index.php).

6.4 CONFORMATION: DESCRIPTION OF SOME POLYSACCHARIDE STRUCTURES

The different kinds of primary structures that result in secondary and tertiary structures give different kinds of properties such as water solubility, aggregation and crystallization, viscosity, gel formation, digestibility and biological recognition. Structural polysaccharides are almost always linear molecules, while polysaccharides that serve primarily as reserve foods are commonly branched or in the case of starch a mixture of linear and branched polysaccharides with the branched type predominating. In general, linear molecules are excellent structural materials because they pack closely and form many intermolecular secondary attachments, which make the structure strong, rigid and insoluble or at least sparingly soluble. Branched polysaccharides, on the other hand, are easily soluble in water and have immense thickening powers. The structural characteristics of some glycans are described below:

6.4.1 Starch

Starch is a polysaccharide composed exclusively of D-glucose and is one of the three most abundant organic compounds found on Earth (cellulose and murein being the other two abundant compounds). Most starches are composed of two types of polysaccharides; amylose and amylopectin. The former is a mixture of linear polysaccharides of D-glucose units linked α-1 → 4 to each other. The latter consists of a mixture of branched polysaccharides of D-glucose unit linked α-1 → 4, with 5% of α-1 → 6 branch linkages corresponding to an average of 20 D-glucose units per branch chain. The average number of glucose residues for amylose can vary from 250 to 5000, and the average number of glucose residues for amylopectin can vary from 10 000 to 100 000. Amylopectin is a much larger molecule than amylose. It may be considered to consist of several amylose chains, one linked to another by α-1 → 6 branch linkages. The branch chains consist of three types:

1. B1-chains that are relatively long and join the individual clusters together;
2. B2-chains that have one or more chains attached to them by branch linkages; and
3. A-chains that are branch chains without any other branching linkages attached to them.

In an aqueous solution, amylose has a random coil structure with a variable amount of single helical structure composed of six, seven or eight glucose residues per turn of helix (Szejtli *et al.*, 1967). When amylose undergoes retrogradation (precipitation from solution), the molecules associate together to form double helices that further associate to give the precipitate.

X-ray diffraction indicates that the starch granules have crystalline properties with varied degree of crystallinity. Waxy starches that are 100% amylopectin shows a crystallinity of 40% and high amylose starches show a crystallinity of only 15%. This suggests that it is the amylopectin component that is primarily involved in the crystalline regions of the starch granule, and the amylose component is in the amorphous region of the granule. It is postulated that the starch granule consists of crystalline regions that are interspersed with amorphous regions. The crystalline regions consist of double helical chains of amylopectin that are ordered by association with each other, and the amorphous regions consist of starch chains that are nonassociated and are hydrated to form a gel. The amorphous regions are less dense, more open, and much more susceptible to acid and enzyme hydrolysis than the crystalline regions. The α-1 → 6 branch linkages of amylopectin are points for promoting double helices that provide the interpretation for the structure of the crystalline regions of the starch granule that are resistant to acid and enzyme hydrolysis. ^{13}C-NMR studies indicate double helical contents of 38–53% for starch granules (Gidley and Bociek, 1988).

6.4.2 Glycogen

Glycogen is a polysaccharide that serves as a form of reserve storage of chemical energy in animals. It is particularly prevalent in liver and skeletal muscle. Liver glycogen is used to maintain a constant level of blood glucose. Skeletal muscle, brain and heart muscle glycogens supply glucose as an immediately available source of chemical energy for the physiological function of these tissues. The brain normally uses about 100 g of glucose from glycogen per day. Liver glycogen is capable of providing 100–150 mg of glucose per min to blood over a sustained period of 12 h if necessary. Glycogen is an α-1 → 4 linked glucan with 10% α-1 → 6 branch linkages. It is highly water-soluble with properties similar to amylopectin. However, glycogen is more highly branched than amylopectin (10% versus 5%) and has an average chain length of 10–12 *versus* an average chain length of 20–23 for amylopectin (Marshall, 1974). Another major difference is that amylopectin occurs as a partial crystalline material in a starch granule, and glycogen is amorphous and does not occur in a large crystalline granule.

Glycogen, like amylopectin, is a polydisperse molecule ranging in size from 1×10^3 to 2×10^6 kDa with the largest amounts at the lower molecular weights (Geddes *et al.*, 1977). The compact structure, low osmotic pressure and ready enzymatic availability of about 40% of the glucose residues of glycogen make it an efficient material as a reserve carbohydrate energy source.

6.4.3 Pectins

Pectins are polysaccharides particularly prevalent in fruits such as apple pulp (10–15%) and orange and lemon rinds (20–30%). When fruit becomes overripe, the pectin is broken down into its constituent monosaccharide sugars. As a result, the fruit becomes soft and loses it firmness. Pectins are composed of D-galactopyranosyl uronic acid units linked α-1 → 4. A relatively large number of the carboxyl groups of the uronic acids exist as methyl esters. Some pectins also have 2-*O*-acetyl or 3-*O*-acetyl group on the D-galactopyrosyl uronic acid units. There are also a small number (1 in 25) of α-L-rhammnopyranosyl units attached to the C-2 position of the uronic acid units. The average molecular weights of pectins are reported to be in a range of 20–400 kDa. Pectins are divided into two categories:

1. pectinic acids, which are polygalacturonic acids that are partially methylated; and
2. pectic acids, which are polygalacturonic acids with no or only a negligible amount of methyl ester.

The pectinic acids can be subdivided into two types: LM-pectins (low methyl pectins) are those that contain less than 50% methyl esters, and HM-pectins (high methyl pectins) are those that contain greater than 50% methyl esters.

One of the most prominent characteristics of pectins is their ability to form gels at concentrations as low as 0.3–0.7% (w/v). Once the gels are formed, the HM-pectins cannot be melted, but the LM-pectins are thermoreversible and can be melted and reformed repeatedly. The rate of gelation is directly proportional to the degree of esterification of the carboxyl group. Rapid-gelling pectins have 70–75% methyl ester, medium-gelling pectins have 65–70% methyl ester, and slow-gelling pectins have 55–65% methyl ester. To form a gel, the pectin molecule must have several areas along its linear structure that will form complexes with several other areas along the structure of other pectin molecules. These areas are called junction zones and must be of a limited size so as to form a gel and not form a precipitate. The substitution of α-L-rhammnopyranosyl units at the 2-O position of the D-galacturonan chain produces discontinuities in the chain and thereby limits the size of the junction zones.

X-ray diffraction studies have shown that pectate, pectic acid and pectinic acid exist as a tight helix, with three D-galacturonic acid residues per turn of the helix (Walkinshaw and Arnott, 1981). Pectinic acid helices pack as parallel chains in which there is a columnar stacking of the methyl ester groups, providing cylindrical hydrophobic areas that are parallel to the helix axis. This produces an aggregation of the chains that are stabilized by hydrophobic interactions of the methyl groups. The structure is also stabilized by hydrogen bonding between hydroxyl groups of residues on different chains.

6.4.4 Cellulose

Cellulose is considered to be the most abundant organic compound on Earth. It is the major structural component of the cell wall of higher plants. It is a major component of cotton boll (~100%), flax (80%), jute (60–70%) and wood (40–50%). Celluloses from all sources are high molecular weight linear polysaccharides of D-glucopyranose units joined together by β-1 → 4 linkage. The β-(1 → 4) glucosidic linkages adopt a fully extended conformation in which the glucose units zig-zag along the polymeric chain. This conformation gives chains that have every other glucose residue rotated 180°, allowing a high propensity to form intermolecular hydrogen bonds. This results in large aggregates of parallel cellulose chains that have crystalline properties.

The tertiary structure of parallel-running intermolecular hydrogen-bonded cellulose chains further associate by hydrogen bonds and van der Waals forces to produce three-dimensional microfibrils. The microfibrils give an X-ray diffraction pattern that indicates a regular, repeating crystalline structure interspersed by less-ordered paracrystalline regions (Hess *et al.*, 1957). The β-1 → 4 glycosidic linkage produces a structure that has low water solubility. The further secondary and tertiary structural effects of cellulose chains to associate and form fibers makes it very water insoluble and thus impermeable to water. This highly associated, water-insoluble aggregate provides a strong matrix for the cell walls of plants. There are several polysaccharides that have structures similar to cellulose with β-1 → 4 linked monosaccharide units. They are the hemicelluloses, chitin, murein, algin, xanthan and related bacterial polysaccharides.

6.4.5 Chitin

The structure of chitin is essentially the structure of cellulose, with the hydroxyl group at C-2 of the D-glucopyranose residue substituted with an *N*-acetylamino group. Chitin is the structural polysaccharide found in fungi, yeast, green algae, and brown and red seaweed cell walls. Chitin is also the major component of the exoskeleton of insects. It is found in the cuticles of annelids, molluses and in the shells of crustaceans such as shrimp, crab and lobster. Chitin forms intermolecular hydrogen bonds to create fibers. This is a highly ordered, crystalline structure. Chitin exists in three types of intermolecular hydrogen-bonded structures, called α-, β- and γ-chitin. The chains in α-chitin are bonded in an antiparallel structure. The chains in β-chitin are parallel, and in γ-chitin the chains are bonded with two parallel chains interspersed by an antiparallel chain (Rudall, 1963). The most abundant form is α-chitin, which also is the most stable. β- and γ-Chitins occur where the properties of the polysaccharide require flexibility and toughness. When chitin is treated with a strong alkali, the *N*-acetyl groups are removed to give a readily water-soluble polysaccharide of β-(1 → 4)-poly-2-amino-2-deoxy-D-glucopyranose, called chitosan. Chitosan has a number of medical uses such as wound dressings, drug delivery agents, hypocholesterolemic agents and for use in contact lenses. It is also used in the purification of drinking water and in cosmetics and personal care products. Chitosan, being a β-glucan, is not digested by humans and can serve as a dietary fiber.

One significant difference between cellulose and chitin is whether the chains are arranged in parallel (all the reducing ends together at one end of a packed bundle and all the nonreducing ends together at other end) or antiparallel (each sheet of chains having the chains arranged oppositely form the sheets above and below). Natural cellulose seems to occur only in parallel arrangements. However, chitin can occur in three forms, sometimes all in the same organism.

6.5 GLYCOBIOLOGY: STUDY OF GLYCOPROTEIN-ASSOCIATED GLYCANS

6.5.1 Glycoprotein and glycoforms

The covalent association between glycans (broadly oligo- and polysaccharides) and proteins or lipids are combined under the collective name glycoconjugates to which belong the proteoglycans and glycoproteins (glycosylated proteins). Figure 6.3 relates various classes of glycoconjugates.

The carbohydrate part of glycosylated proteins can consist of one, several or many residues. Artificial glycosylated proteins used as model compounds for experimental purposes are called neoglycoproteins. Peptidoglycans are glycosaminoglycans that are cross-linked with peptides. Proteoglycans are a sub-class of the glycoproteins, in which the carbohydrate residues are glycosaminoglycans (generally in disaccharide repeat units) in linear chains. A proteoglycan can carry up to 100 glycosaminoglycan chains with up to 200 disaccharide repeat units. Glycopeptides contain carbohydrates, mostly oligosaccharides, which are bound to oligopeptides. Glycoproteins contain carbohydrates bound to proteins via glycosidic linkages (Corfield, 2000). The carbohydrates can be monosaccharides, oligosaccharides or polysaccharides and also their derivatives. In many glycoproteins (from serum to membranes) the carbohydrates are found as oligosaccharides, either linear or branched. The highly branched oligosaccharides can be composed of up to 20 monosaccharide residues building up from repeating units. Studies of the structure and function of the carbohydrates in glycosylated proteins is known as glycobiology, which deals with the role of carbohydrates in biological systems (Rademacher *et al.*, 1988; Varki, *et al.*, 1999; Bertozzi and Kieswsling, 2001).

Glycoproteins are essential to many basic cellular and disease processes including molecular/cellular recognition, intracellular sorting, cell growth, fertilization, immune defense, inflammation, tumor metastasis, viral replication and bacterial/parasite infection. For example, carbohydrates of glycoproteins at the mammalian cell surface have two major functions. Inside the cell, they help proteins fold and assemble correctly in the

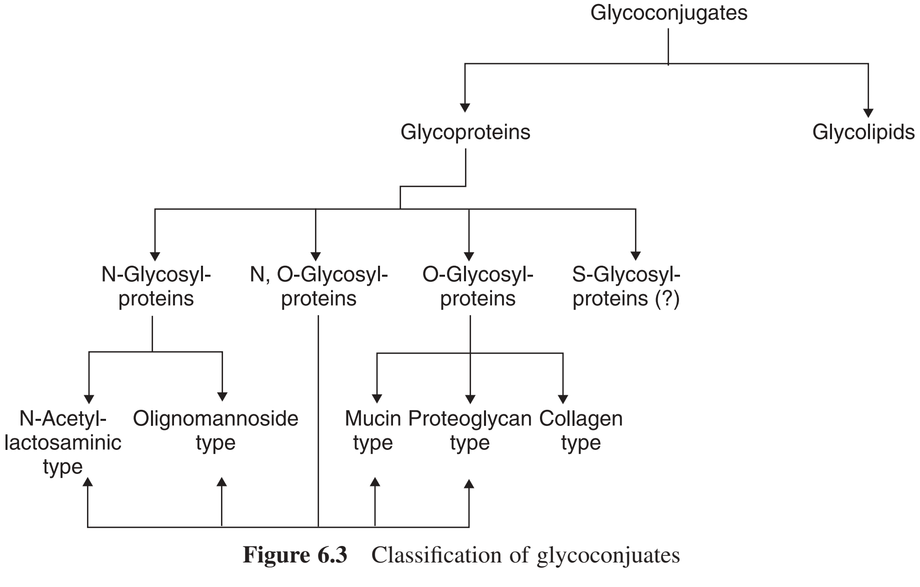


Figure 6.3 Classification of glycoconjuates

endoplasmic reticulum and act as a signal for the correct migration of glycoproteins. Outside the cell, they provide specific recognition structure for interaction with a variety of external ligands. The carbohydrate groups confer important properties to proteins such as conformational stability, protease resistance, charge and water binding capacity, specificities in biological recognition, signals for targeting and molecular/cellular recognition. The post-translational glycosylation of polypeptides to form glycoproteins has the following characteristics:

- Different glycoproteins from the same cell may contain different oligosaccharide structures.
- An individual glycoprotein may contain multiple glycosylation sites.
- An individual polypeptide usually carries several different oligosaccharide structures, many of which are found at the same glycosylation site referred to as site heterogeneity (microheterogeneity).
- The oligosaccharide heterogeneity at a single glycosylation site is reproducible during a constant physiological condition.
- The oligosaccharide heterogeneity results in a set of glycosylated structures known as glycoforms with different physical and biochemical properties that may lead to functional diversity.
- The oligosaccharide processing is cell and tissue specific.

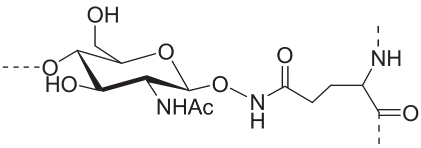
The microheterogeniety, i.e. consequence of multiple glycosylation sites and partial site-occupancy of a population of different oligosaccharide structures, results in a set of glycosylated variants of a common polypeptide. These glycoforms differ in the nature, number, location and sequence of oligosaccharides, which affect their properties and functions.

6.5.2 Structure diversity of oligosaccharide chains

The major glycoses in glycoproteins consist of D-galactose (Gal), D-galactosamine/*N*-acetyl-D-galactosamin (GalN/GalNAc), D-glucose (Glc), D-glucosamine/*N*-acetyl-D-glucosamine (GlcN/ClcNAc), L-fucose (Fuc), D-mannose (Man), D-xylose (Xyl), Neuraminic acid (Neu), and *N*-acetylneuraminic acid (sialic acid, Sia)/*N*-glycolylneuraminic acid (NeuAc/NeuGc). These glycoses are attached to polypeptides in one of three ways:

6.5.2.1 N-linked oligosaccharide chain. An N-glycosidic bond to the side-chain of asparagine.

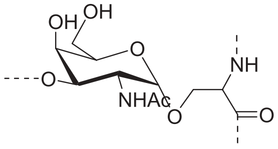
N-linked glycose chains contain an GlcNAc residue at their reducing termini and are linked

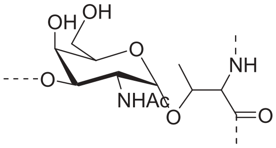


to the amide group of an Asn of proteins where asparagine must be present in the subsequence, Asn-X-Ser/Thr (X can be any amino acid except Pro).

N-linked Man-rich oligosaccharides are synthesized on a specific lipid (dolicholphosphate) then added to protein molecules prior to their translocation. The microheterogeneity (a range from 10 to 30 different oligosaccharide structures can be present at each site) from the trimming and addition of new glycoses (GlcNAc and Gal) to the oligosaccharide chains. They act as signals of cell surface recognition processes.

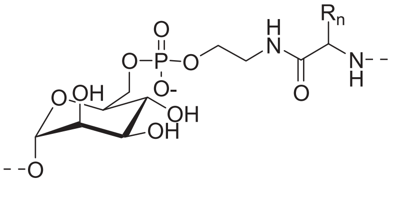
6.5.2.2 O-linked oligosaccharide chain. An O-glycosidic bond to the side-chain of serine, threonine.





O-linked glycose chains contain an GalNAc residue at their reducing termini and are linked to the hydroxyl group of Ser or Thr. A few O-linkages to hydroxylysine or hydroxyproline have been observed (e.g. collagens). O-linked chains are built up by the direct addition of glycoses to the proteins. They confer particular physicochemical properties on glycoproteins. Some oligosaccharide chains are attached to cysteine residue in S-linked glycoproteins.

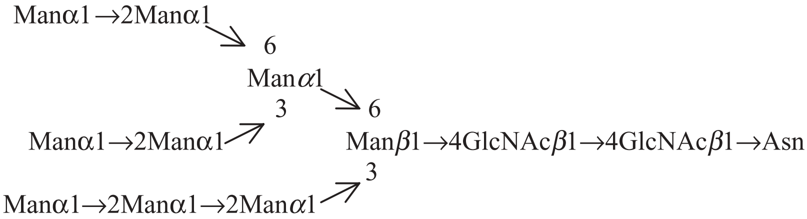
6.5.2.3 As part of the glycosylphosphatidylinositol membrane anchor. The glycosylphosphotidylinositol (GPI) anchors a wide variety of proteins, which serve as transmembrane polypeptide domains of the eukaryotic plasmic membrane. The phosophoinositol glycolipid attachment occurs by the formation of a C-terminal amide link to ethanolamine, which is connected to glycan (i.e. 6-hydroxyl of α-D-mannose) via a phosphodiester bond resulting in the attachment of the modified protein to a cell membrane, i.e. protein–CONH–$(CH)_2OP(O)_2O$–glycan–*O*–inositol phospholipid.



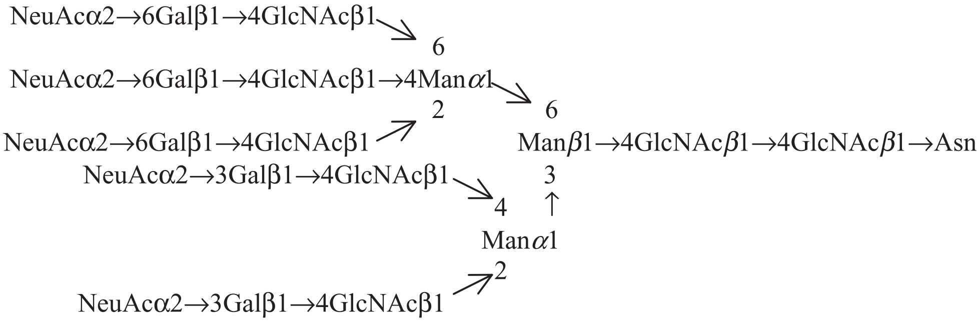
Six amino acids, namely Cys, Asp, Asn, Gly, Ala and Ser (CDNGAS), serve as a GPI attachment site. This attachment is common for the surface coat-protein of certain parasites and membrane enzymes.

In addition to the glycosidic-peptidic linkages, N- and O-linked oligosaccharide chains differ in their core structures, while peripheral sequences exhibit similarities, especially in the longer sugar chains. All N-linked glycose chains contain the pentasaccharide, Manα1 → 6(Manα1 → 3)Manβ1 → 4GlcNAcβ1 → 4GlcNAc as a common core (trimannosyl core) and are classified into three major subgroups depending on the glycose extensions to the trimannosyl core (Kobata, 1993) as:

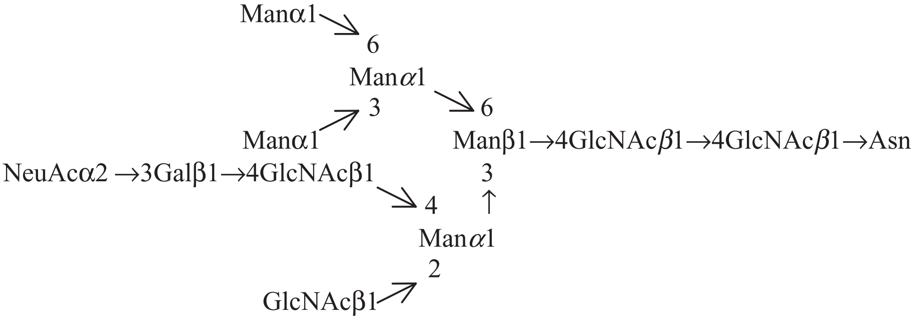
(A) High mannose (oligomannose) type oligosacchride contains only α-mannosyl residues in addition to the core and serves as a biosynthetic precursor of the two complex types. Heptasaccharide (without Manα1 → 2) is also included in this type.


Manα1→2Manα1
6
Manα1
3
6
Manα1→2Manα1
Manβ1→4GlcNAcβ1→4GlcNAcβ1→Asn
3
Manα1→2Manα1→2Manα1

(B) Complex type oligosaccharide contains no mannose residues other than those in the core. The presence or absence of αFuc to the C-6 position of the proximal GlcNAc and the βGlcNAc to the C-4 position of the βMan of the trimannosyl core contributes to the structural variations.


NeuAcα2→6Galβ1→4GlcNAcβ1
6
NeuAcα2→6Galβ1→4GlcNAcβ1→4Manα1
2
6
NeuAcα2→6Galβ1→4GlcNAcβ1
Manβ1→4GlcNAcβ1→4GlcNAcβ1→Asn
NeuAcα2→3Galβ1→4GlcNAcβ1
3
4
Manα1
2
NeuAcα2→3Galβ1→4GlcNAcβ1

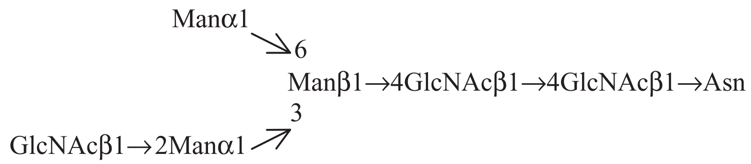
(C) Hybrid type contains the characteristic features of both complex type (NeuAcα2→ 3Galβ1→4GlcNAcβ) and the mannose type glycose chains. The presence or absence of αFuc and βGlcNAc again offers structural variations.


Manα1
6
Manα1
3
6
Manα1
Manβ1→4GlcNAcβ1→4GlcNAcβ1→Asn
NeuAcα2 →3Galβ1→4GlcNAcβ1
3
4
Manα1
2
GlcNAcβ1

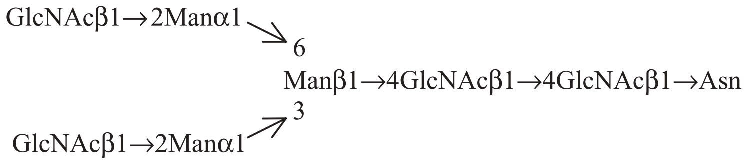
The trimannosyl core common to all N-linked glycose chains may carry branches, (GlcNAcβ1→4) and (Fucα1→6) at βMan and βGlcNAc(-Asn) respectively. In addition, poly-*N*-aceyllactosamine type contain repeating units of (Galβ1→4GlcNAcβ1→3) attached to the core. The lactosamine repeats are not uniformly distributed and may be branched.

The complex type has the largest structural variations, termed antennary, caused by the various branchings of outer chains from the trimannosyl core. Thus one to five branches joined to the trimannosyl core result in the formation of mono-, bi-, tri-, tetra- and pentaantennary glycans:

Monoantennary:

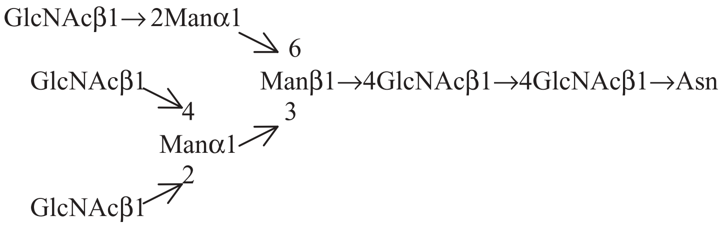

Manα1
6
Manβ1→4GlcNAcβ1→4GlcNAcβ1→Asn
3
GlcNAcβ1→2Manα1

Biantennary:

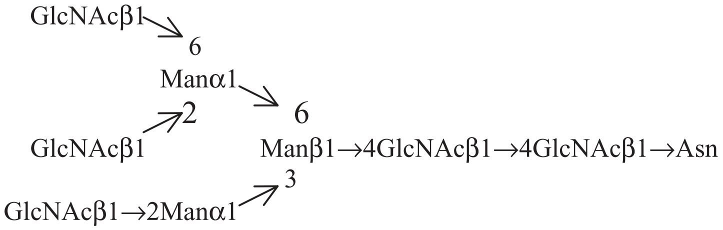

GlcNAcβ1→2Manα1
6
Manβ1→4GlcNAcβ1→4GlcNAcβ1→Asn
3
GlcNAcβ1→2Manα1

Triantennary:

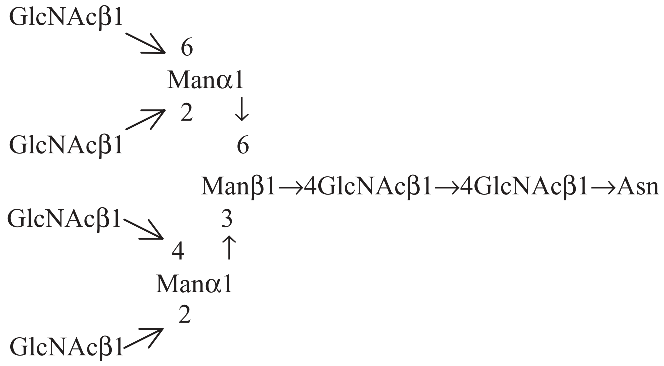
2,4-branched


GlcNAcβ1→2Manα1
6
GlcNAcβ1
4
Manβ1→4GlcNAcβ1→4GlcNAcβ1→Asn
3
Manα1
2
GlcNAcβ1

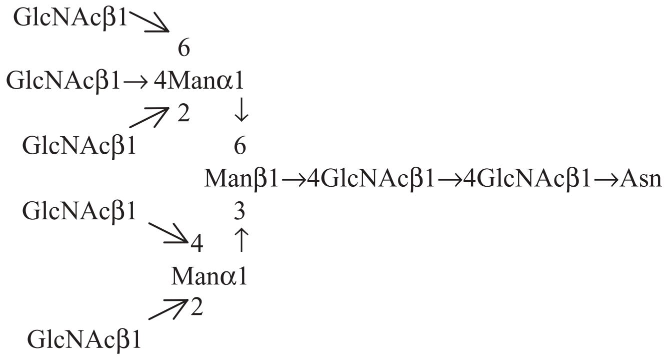
2,6-branched


GlcNAcβ1
6
Manα1
2
6
GlcNAcβ1
Manβ1→4GlcNAcβ1→4GlcNAcβ1→Asn
3
GlcNAcβ1→2Manα1

Tetraantennary:

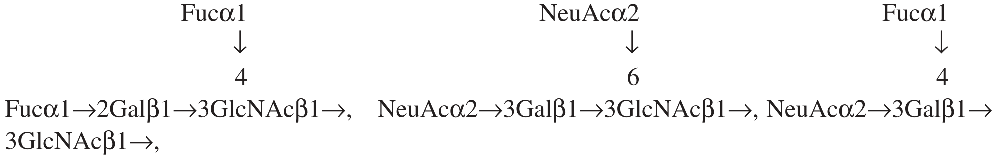

GlcNAcβ1
6
Manα1
2
6
GlcNAcβ1
Manβ1→4GlcNAcβ1→4GlcNAcβ1→Asn
GlcNAcβ1
3
4
Manα1
2
GlcNAcβ1

Pentaantennary:


GlcNAcβ1
6
GlcNAcβ1→4Manα1
2
GlcNAcβ1
6
Manβ1→4GlcNAcβ1→4GlcNAcβ1→Asn
GlcNAcβ1
3
4
Manα1
2
GlcNAcβ1

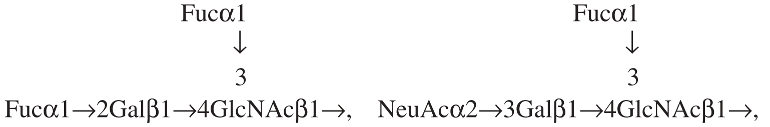
Various outer chains extend the antennary structures. Some of the common outer chains are:

Galβ1→3GlcNAcβ1→, Fucα1→2Galβ1→3GlcNAcβ1→,



$$\text{Fuc}\alpha1\rightarrow2\text{Gal}\beta1\rightarrow3\overset{\substack{\text{Fuc}\alpha1\\ \downarrow\\ 4}}{\text{GlcNAc}}\beta1\rightarrow,\quad \text{NeuAc}\alpha2\rightarrow3\text{Gal}\beta1\rightarrow3\overset{\substack{\text{NeuAc}\alpha2\\ \downarrow\\ 6}}{\text{GlcNAc}}\beta1\rightarrow,\quad \text{NeuAc}\alpha2\rightarrow3\text{Gal}\beta1\rightarrow3\overset{\substack{\text{Fuc}\alpha1\\ \downarrow\\ 4}}{\text{GlcNAc}}\beta1\rightarrow,$$

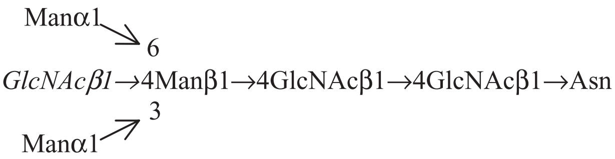
Galβ1→4GlcNAcβ1→, Fucα1→2Galβ1→4GlcNAcβ1→, NeuAcα2→3(6)Galβ1→4GlcNAcβ1→,



$$\text{Fuc}\alpha1\rightarrow2\text{Gal}\beta1\rightarrow4\overset{\substack{\text{Fuc}\alpha1\\ \downarrow\\ 3}}{\text{GlcNAc}}\beta1\rightarrow,\quad \text{NeuAc}\alpha2\rightarrow3\text{Gal}\beta1\rightarrow4\overset{\substack{\text{Fuc}\alpha1\\ \downarrow\\ 3}}{\text{GlcNAc}}\beta1\rightarrow,$$

Galα1→3Galβ1→4GlcNAcβ1→, SO_4-4GalNAcβ1→4GlcNAcβ1→.

The combination of different antennary structures and various outer chains produces a large number of different complex type N-linked oligosaccharide chains of glycoproteins. Thus, the antennae of mature glycans usually consist of one or more *N*-GlcNAc units with the chains terminating in either NeuAc (sialic acid) or Gal. Fuc is frequently attached to the Asn-linked GlcNAc and often additionally on the antennae. Other common modifications to the basic structure include a *GlcNAc* residue (italic) attached to the 4-position of the core branching Man residues, referred to as a 'bisecting' GlcNAc residue:



$$\begin{array}{l}\text{Man}\alpha1\searrow{}^{6}\\ \textit{GlcNAc}\beta\textit{1}\rightarrow4\text{Man}\beta1\rightarrow4\text{GlcNAc}\beta1\rightarrow4\text{GlcNAc}\beta1\rightarrow\text{Asn}\\ \text{Man}\alpha1\nearrow{}^{3}\end{array}$$

and sulfate groups, which can be found in a variety of locations.

The N-linked glycans of glycoproteins can be functionally divided into two types: intra- and extracellular. Intracellularly, the broad function of N-linked glycans is protein trafficking (Helenius and Aebi, 2001) such as a folding sensor of protein folding in the calnexin-calreticulin cycle. Extracellularly, N-linked glycans can function as structural elements and as ligands for receptors. Structurally, the large and flexible N-linked glycans can increase protein stability by restricting the conformational flexibility of the underlying protein without sacrificing net entropy of the system. For example, in human CD2, the protein-proximal core GlcNAc-GlcNAc is in close contact with a cluster of charged and polar residues located on the face of a β-sheet. The glycan acts in concert with the polypeptide to orchestrate the overall structure and function of the glycoprotein. N-linked glycans also function extracellularly as ligands for carbohydrate receptors. For example, variation in the carbohydrate content of the N-linked glycans of erythropoietin alters the serum half-life of the glycoprotein and changes its *in vivo* activity to stimulate red cell proliferation and differentiation in bone marrow (Fukuda *et al.*, 1989).

The oligosaccharide chains of O-linked glycans are generally simpler and less branched than N-linked glycans. O-linked-saccharides can simply be single sugar residues, or complex structures built on an initial GalNAc, Fuc, or Man that is attached to the Sere/Thr residue of polypeptides. O-linked glycans do not share a common core structure.

So far six types of core structures (with initial GalNAc) are recognized in mucin-type O-linked glycans known as Tn antigens:

Core 1: Galβ1→3GalNAcα1→Ser/Thr
Core 2: GlcNAcβ1↘
6
GalNAcα1→Ser/Thr
3
Galβ1↗
Core 3: GlcNAcβ1→3GalNAcα1→Ser/Thr
Core 4: GlcNAcβ1↘
6
GalNAcα1→Ser/Thr
3
GlcNAcβ1↗
Core 5: GlcNAcβ1→6GalNAcα1→Ser/Thr
Core 6: GalNAcβ1→3GalNAcα1→Ser/Thr

Thus the only common structural element in these O-linked oligosaccharide chains is GlcNAc from which various outer chains are attached such as:

Linear: Galβ1→3, GlcNAcβ1→6, NeuAcα2→6, Fucα1→2Galβ1→3GlcNacβ1→3,
Branch: NeuAcα2→3Galβ1→3 and NeuAcα2→6,
Galβ1→4GlcNAcβ1→3 and Fucα1→3(NeuAcα2→3Galβ1→4)GlcNacβ1→6,
Fucα1→2(GalNAcα1→3)Galβ1→4GlcNAcβ1→3Galβ1→4GlcNAcβ1→3Galβ1→3 and Fucα1→2Galβ1→4GlcNAcβ1→6.

Mucin-type glycosylation is found on many cell-surface proteins and plays a vital role in interactions of the cell with its environment, particularly for immune responses (Tsuboi and Fukuda, 2001).

The tetrasaccharide core of GPI consists of Manα1→2Manα1→6Manα1→4GlcNAc, which is anchored to *myo*-inositol-1-phosphate of the membrane. The membrane-associated lipid is commonly composed of saturated 14- to 18-C esters/ethers to glycerol. In mammalian cells, GPI-anchored proteins are involved in fundamental processes such as signal transduction, cell membrane processing and homeostasis and immune responses, as well as the pathobiology of cancers and infectious diseases (Nosjean *et al.*, 1997). GPI-linked proteins freely diffuse in the lipid membrane, but many are found on mammalian cells associated with other membrane-bound proteins in cholesterol and sphingolipid enriched microdomains (termed lipid rafts). Raft-associated GPI proteins are essential factors for many cellular phenomena, such as immune synapse formation, virus/host infection and the pathogenesis of human prion, a GPI protein (Kaneko *et al.*, 1997).

Typically, the outer chains of glycoproteins are often short utilizing galactose, fucose, N-acetylglucosamine, *N*-acetylgalactosamine, and *N*-acetylneuraminic acid (sialic acid), but both *N*-linked and O-linked oligosaccharides can reached sizes of 15–40 residues or more. The choice and usage of these glycoses joined by different linkages at different locations in the sequence give rise to enormous structural diversity of the glycoprotein-associated glycans. Glycan information (including structure) of glycoproteins and other complex carbohydrates can be retrieved from CarbBank at http://bssv01.lanes.ac.uk/gig/pages/gag/carbbank.htm or Complex Carbohydrate Structure Data of SugaBase at http://www.boc.chem.uu.nl/sugabase/database.html. The Sweet

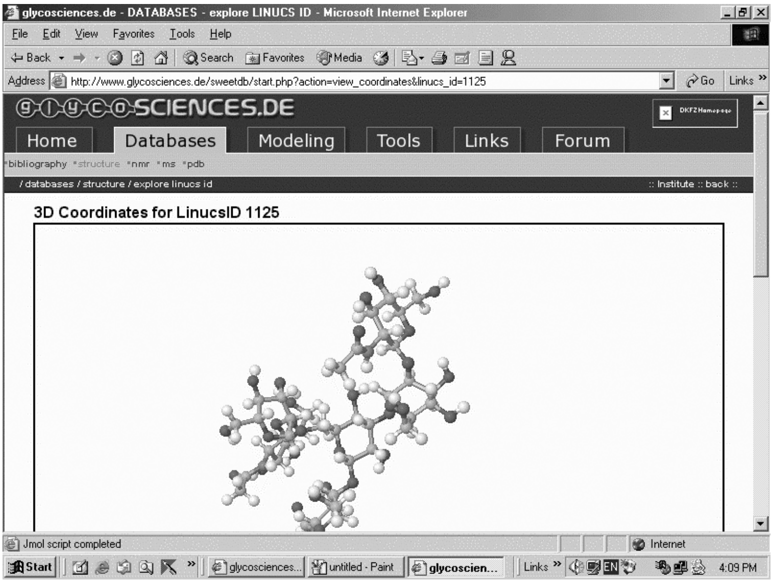

glycosciences.de - DATABASES - explore LINUCS ID - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.glycosciences.de/sweetdb/start.php?action=view_coordinates&linucs_id=1125
Go Links
GLYCOSCIENCES.DE
DKFZ Homepage
Home Databases Modeling Tools Links Forum
*bibliography *structure *nmr *ms *pdb
/ databases / structure / explore linucs id
:: Institute :: back ::
3D Coordinates for LinucsID 1125
Jmol script completed
Internet
Start
glycosciences...
untitled - Paint
glycoscien...
Links
4:09 PM

Figure 6.4 Retrieval of glycan structure from database of Glycosciences
Glycan structures can be searched and retrieved from database of Glycosciences (http://www.glycosciences.de/sweetdb/index.php). For example entering Hex = 3, dHex = 1 and HexNAc = 4 in the composition search option, the query returns a list of hits (listing LinucsID, composition, formula, and source of glycans) from which the desired glycan is selected for viewing molecular structure and relevant information. Clicking "Explore" then "theor. 3D Co-ord" displays the 3D structure of the glycan and its atomic coordinate (pdb format) can be downloaded.

database of Glycosciences at http://www.glycosciences.de/sweetdb/index.php provides facilities for retrieving and displaying glycan structures (Figure 6.4).

6.5.3 Structural analysis

The analytical strategies for the structural studies of glycoproteins (Dwek *et al.*, 1993) involve:

- Information on the presence of *N*- or O-glycosylation, extent of glycosylation and monosaccharide composition of the glycan.
- Separation of individual glycosylated sites generally as glycopeptides, quantification of the degree of occupancy/glycosylation, and amino acid sequence of the glycopeptides (glycosylated sites).
- Release of oligosaccharides from the glycopeptides and purification of the oligosaccharides.
- Determination of the monosaccharide sequence of each oligosaccharide.

The procedures generally include:

6.5.3.1 Identification of glycosylated sites. A glycosylated site is generally determined by isolating the appropriate glycopeptide after proteolysis of the glycoprotein. The prediction of glycosylation sites of proteins can be performed at NetOGlyc server (http://www.cbs.dtu.dk/services/NetOGlyc/). The amino acid analyses of the glycopetide are carried out before and after the treatment with endoglycosidases/endoglycopeptidases to assign the peptide, which can then be isolated for carbohydrate analysis. Glycopeptides isolated are almost invariably heterogeneous with respect to their attached glycans and the full complement of glycan heterogeneity is often recovered in a single glycopeptide pool. Carbohydrate analysis can be performed after the release of glycans directly from the glycoprotein or glycopeptide.

6.5.3.2 *Release of glycans from glycopeptide/glycoproteins.* The determination of the structures of glycans of glycoproteins/glycopeptides begins with their release from the peptide/protein attachments. Oligosaccharides attached to Ser and Thr (O-linked) or Cys (S-linked) can be released by a β-elimination using a mild alkaline (0.05–0.5 N NaOH) treatment. The reaction is usually conducted in the presence of sodium borohydride (0.15–1.0 M), which reduces the reducing end of the released oligosaccharide. This reduction prevents alkaline degradation of the oligosaccharide by the peeling reaction, and it yields information about the carbohydrate residue(s) that is/are attached to the protein/peptide as alditol(s). A standard procedure uses 0.1 N NaOH and 0.3 M $NaBH_4$ at 37°C for 48 h. (Fukuda, 1989).

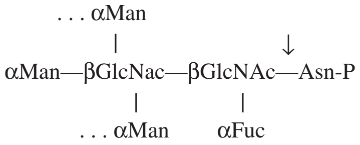
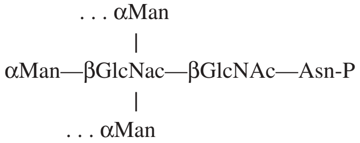
Both N- and O-linked glycans can be released from glycoproteins with hydrazine (Takasaki, *et al.*, 1982). The N-linked oligosaccharides can be cleaved by hydrazinolysis (Takasaki *et al.*, 1982; Patel and Parekh, 1994). The glycoprotein/glycopeptide is heated at 95°C with anhydrous hydrazine for 4 h. The procedure is conducted by suspending 0.2–1 mg of salt-free, freeze-dried glycoprotein/glycopeptide in 0.5–1.0 mL of freshly distilled anhydrous hydrazine. The solution is heated in a sealed tube at 95°C for 4 h. The glycoprotein/glycopeptide sample usually dissolved after 1 h. To release both *O*- and *N*-glycosidic bonds, the glycoprotein/glycopeptide (0.2–1 mg) is heated at 100°C with anhydrous hydrazine (0.5–1.0 mL) for 8–12 h. For the sequential release of O- and N-linked oligosaccharides (Patel and Parekh, 1994), the hydrazinolysis is performed first at 60°C for 5 h. (release O-linked oligosaccharides) and then at 95°C for 4 h. (release N-linked oligosaccharides). Hydrozenolysis releases glycans that are intact with free reducing ends and is nonselective with respect to the glycan but the peptide bonds are destroyed. Although O-linked glycans are sometimes released from glycoproteins with hydrazine, they are more usually released by β-elimination with alkali. Because the resulting reducing sugars are unstable at high pH, a reducing agent such as sodium borohydride is added to reduce them to the alditols.

Another chemical method that liberates both O- and N-linked oligosaccharides from glycoproteins involves anhydrous trifluoromethane sulfonic acid or anhydrous hydrogen fluoride. Trifluoromethane sulfonate hydrolysis is performed at 0°C for 0.5–2 h under nitrogen. The reaction mixture is cooled below −20°C in a dry ice-ethanol bath and slowly neutralized with 60% (v/v) aqueous pyridine (previously cooled to −20°C).

Various endoglycosidases, which cleaves oligosaccharide chains from glycoproteins, are commonly employed to liberate oligosaccharides from glycopeptides (glycosylated sites) and glycoproteins (Table 6.8). For example, aminohydrolases, peptide-N^4-(*N*-acetyl-β-glycosaminyl)asparagine amidases (PNGases) specifically release oligosaccharide chains from N-linked glycopeptides/glycoproteins (Tarentino and Plummer, 1994).

6.5.3.3 *Labeling of the released glycans.* To facilitate their detection in the subsequent procedures, the released oligosaccharides with free reducing ends are labeled. Two methods that are commonly used to label oligosaccharides are reductive amination with a fluorescent compound, such as 2-aminobenzamide and reduction with alkaline NaB^3H_4.

TABLE 6.8 Endoglycosidases which liberate oligosaccharide chains from glycopeptides/glycoproteins

Enzyme	Source	Substrate	Specificity (bond of cleavage at ↓)
Peptide-*N*-glycosidase A (PNGase A), EC 3.5.1.52	almond	All types of N-linked glycopeptides/glycoproteins with/without α1,3/6-bound Fuc	 $\begin{array}{ccccccc} & & \ldots\alpha\text{Man} & & & & \\ & & \vert & & & & \downarrow \\ \alpha\text{Man} & — & \beta\text{GlcNac} & — & \beta\text{GlcNAc} & — & \text{Asn-P} \\ & & \vert & & \vert & & \\ & & \ldots\alpha\text{Man} & & \alpha\text{Fuc} & & \end{array}$
Peptide-*N*-glycosidase F (PNGase F), EC 3.5.1.52	*Flavobacterium Meningosepticum*	All types of N-linked glycopeptides/glycoproteins without α1,3/6-bound Fuc	 $\begin{array}{ccccccc} & & \ldots\alpha\text{Man} & & & & \\ & & \vert & & & & \\ \alpha\text{Man} & — & \beta\text{GlcNac} & — & \beta\text{GlcNAc} & — & \text{Asn-P} \\ & & \vert & & & & \\ & & \ldots\alpha\text{Man} & & & & \end{array}$
Peptide-*O*-glycosidase EC 3.2.1.97	*Diplococcus pneumoniae*	O-linked glycopeptides/glycoproteins	$\beta\text{Gal—}\beta\text{GalNacl}\overset{\downarrow}{\text{—}}\text{Ser/Thr-P}$
Endoglycosidase H (Endo H) EC 3.5.1.26	*Streptomyces plicatus*	High mannose and hybrid types of N-linked glycopeptides/glycoproteins	$\ldots\alpha\text{Man—}\beta\text{GlcNac}\overset{\downarrow}{\text{—}}\beta\text{GlcNAc—Asn-P}$

Notes: 1. The linkages common to N- and O-linked oligosaccharides are not expressed for simplicity.
2. P represents glycopeptide/glycoprotein.
3. Endo H cleaves the glycosidic bond, GlcNAcβ(1→4)GlcNAc leaving the marker (GlcNAc) attached to Asn and can be used for the identification of the glycosylated sites.

6.5.3.4 Purification of released oligosaccharides. Numerous methods have been employed to fractionate/purify the heterogeneous oligosaccharide mixture released from glycoproteins, including various chromatographic and electrophoretic techniques (Hicks, 1988). Sometimes oligosaccharides are derivatized to facilitate fractionation, by gel electrophoresis (Stack and Sullivan, 1992). Often a combination of several techniques must be used to purify an oligosaccharide mixture to homogeneity. Approaches to these chromatographic and electrophoretic techniques will be described in Chapter 8.

6.5.3.5 Determination of monosaccharide sequences. Similar approaches are applied to the monosaccharide sequence analyses of oligosaccharides released from glycoproteins and polysaccharides, except that the structures of oligosaccharides from glycoproteins are more complex by the variation in number and structures of glycoses involved in the chain, their heterolinkages and branch differentials. Usually the combined use of several methods including enzymatic, MS and NMR analyses are needed to determine monosaccharide sequences of complex oligosaccharides (Jiménez-Barbero and Peters, 2003). In practice, glycan structure analysis is facilitated by making certain reasonable assumptions usually supported by monosaccharide composition analysis of the oligosaccharide sample. For example, it can be assumed that all mononsaccharides of mammalian origin have the D-configuration except fucose, which is L, that an N-linked glycan contains the trimannosyl chitobiose core, and that major monosacchrides in glycoproteins consist of Gal, GalNAc, Glc, GlcNAc Fuc, Man and NeuNAc. Approaches that use enzymatic analysis in combination with chromatographic, mass spectrometric and nuclear magnetic resonance techniques have been developed:

Sequential analysis. The labeled oligosaccharide is exposed to highly purified exoglycosidases. After each incubation, the labeled glycan product is identified and the loss of monosaccharide (if any) is determined and measured. This iterative process is continued until no further useful information can be obtained using exoglycosidases.

Immobilization approach. The glycan sample is immobilized to a solid support. The immobilized glycan is then exposed to exoglycosidases sequentially and the monosaccharides released by each enzymatic digestion is detected, identified and quantified.

Parallel analysis: reagent array analysis method. The reagent array analysis method (RAAM) involves dividing a glycan solution into equal aliquots depending on the number of exoglycosidases used in the array, incubating each aliquot with a precisely defined mixture of exoglycosidases, recombining the products of each incubation, and then performing analysis on the pool of products. In essence, a mixture of exoglycosidases is used to digest the sample glycan until a linkage is reached that is resistant to all the exoglycosidases present in that mix. By omitting one or more different exoglycosidase(s) from each mixture, different 'stop point' fragments of the oligosaccharide are generated. By labeling the original oligosaccharide at the reducing terminus, fragments retaining the original reducing terminus are readily distinguished from released monosaccharides. Both positive data (the exoglycosidases hydrolyze linkages up to the stop point) and negative data (the exoglycosidases fail to hydrolyze linkages beyond the stop point) are utilized in the analysis. Chromatographic separation of the combined stop point fragments generates a pattern that is a 'signature' of that glycan treated by the enzyme array used in the analysis. This signature (fingerprint) can be matched against a computer-generated database of theoretical fingerprints to identify the glycan structure. (Dwek, 1996).

6.6 NEOGLYCOPROTEINS

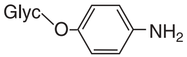
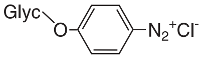
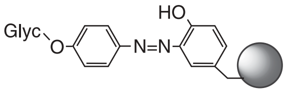
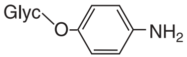
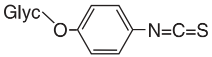
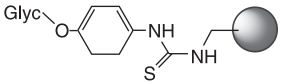
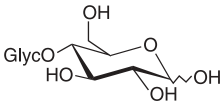
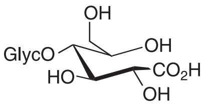
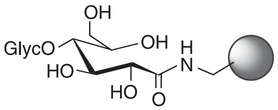
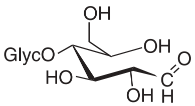
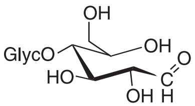
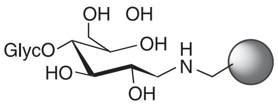
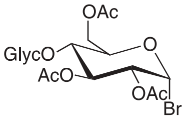
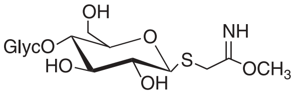
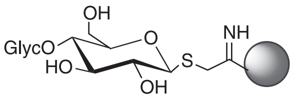
Neoglycoproteins (Stowell and Lee, 1980) refer to synthetic glycoproteins by chemically attaching mono- or oligosaccharides to proteins directly or via spacers. Table 6.9 lists examples of neoglycoproteins and their synthetic methods.

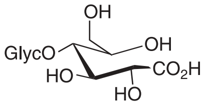
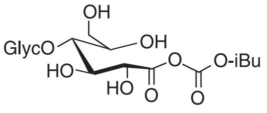
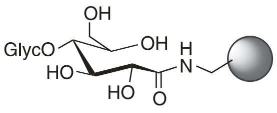
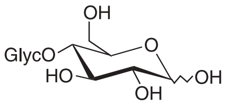
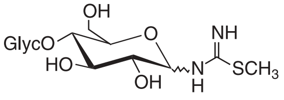
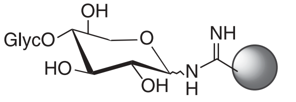
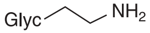
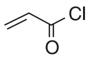
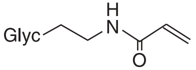
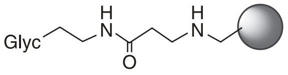
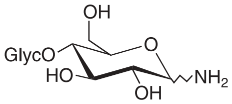
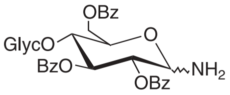
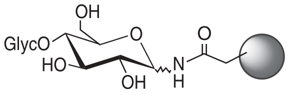
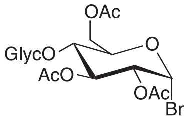
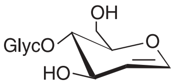
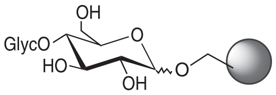
Neoglycoproteins are used to investigate the contribution of glycoses/glycans to the functions mediated by glycoproteins. Given the role of carbohydrate residues as molecular recognition/immunodominant structural elements, the aim is to prepare neoglycoconjugates with higher affinity and better specificity. Applications of neoglyconjugates include specific drug delivery vectors, anti-inflammatory and anti-adhesion agents, glycoporbes and glycosupports, immunodiagnostics as well as immunomodulators.

6.7 ORGANIZATIONAL LEVELS OF BIOMACROMOLECULAR STRUCTURES

The structures of biomacromolecules are organized in distinct hierarchical levels; namely primary, secondary, tertiary, quaternary and quinternary structures (Table 6.10). The polymers of nucleotides, amino acids and glycoses are joined together by phosphodiester, amide and acetal linkages respectively to form nucleic acids, proteins and glycans. They are linear in nucleic acids and proteins whereas both linear and branch chains are possible in glycans. The major repeating structures of all biomacromolecular chains are helix and strand/ribbon. The main helical structures are right-handed helices. They differ in the

TABLE 6.9 Examples of neoglycoproteins and their preparations

	Derivatization			
Preparation	Initial glycose reactant	Reagents	Condensation with protein	Neoglycoprotein
Diazo coupling		HCl,NaNO$_2$/NaHCO$_3$		
Isothiocyanation		CS(Cl)$_2$/pH8–9		
Amidation		I$_2$(NaOH)/pH3–5		
Amination		NaCNBH$_3$, pH7		
Amidination		CS(NH$_2$)$_2$, ClCH$_2$CN/ NaOCH$_3$		

Anhydride reaction		IBuOCOCl, iBu$_3$N/ ~pH10		
guanidination		MeSC(NH$_2$) =NH,OH$^-$/ pH8.5		
Michael addition		 /Et$_3$N		
Carbodiimide condensation		RN=C=NR′; H$_2$(Pd/C)		
Glycal activation		Zn/HOAc; NaOMe		

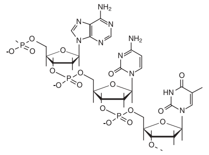
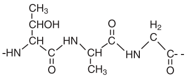
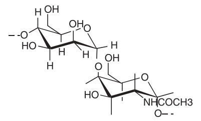
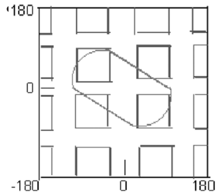
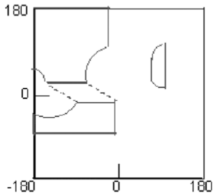
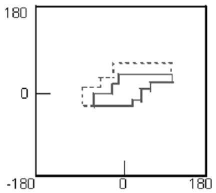
Notes: 1. Only major reagents for derivatization of mono-/ologosaccharide chain (Glyc) for condensation with peptides/proteins are given.

2. The hydroxy groups of Glyc can be protected (Table 8.2), for example with Bz (—CH$_2$—C$_6$H$_5$), which can be removed with either hydrogenation (Pd/C for Bz) or 2,3-dichloro-5,6-dicyano-1,4-benzoquinone (DDQ for p-methoxy-Bz).

3. The spacer arms between Glyc and peptide/protein can be attached by reacting with H$_2$N(CH$_2$)$_2$NH$_2$, HS(CH$_2$)$_2$NH$_2$, CH$_2$=CHCOCl or CH$_2$=CHCONH$_2$.

4. Taken from Stowell and Lee (1980); Romanowsk *et al*. (1994); Sears and Wong (2001).

TABLE 6.10 Different levels of structural organizations of biomacromolecules

	Nucleic acids	Proteins	Glycans
Primary structure	Nucleotide sequence, linear diphosphoesteric linkages, 5′-hydroxy → 3′-hydroxy.  e.g. . . . ACT . . .	Amino acid sequence, linear amide linkages, amino → carboxyl  e.g. . . . TAG . . .	Monosaccharide sequence, linear/branched, glycosidic linkages, nonreduced → reduced  e.g. . . . Ma4ANa . . .
Conformational map (ϕψ plot)			
Secondary Helix	Duplex helix with base pairings, #bases/turn ≥ 10 e.g. 10 bases/turn in B-DNA.	Simplex helix, # residues/turn ≤ 6 e.g. 3.6 residue/turn for α-helix.	Varied (mostly single and double) helix, #glycoses/turn = 5–10 e.g. 6 residues/turn for amylose.
Strand	Antiparallel, paired stem	Parallel and antiparallel, pleated.	Parallel and antiparallel Fibril/ribbon coil
Others	Coil, bulge	Coil, bulge	
Tertiary	Folded conformation e.g. supercoiled DNA, L-structure of tRNA	Domain, over-all fold e.g. all α structure of bacteriorhodopsin, mainly β structure of trypsin, α/β structure of triose isomerase, α+β structure of lysozyme	Folding e.g. crystalline region (double helical fold) of starch, microfibril (parallel chains) of cellulose
Quaternary	e.g. association of rRNAs in ribosome.	e.g. dimeric liver alcohol dehydrogenase, tetrameric hemoglobin	e.g. association of cellulose fibers in cell wall
Quinternary	Nucleoproteins	Nucleo-, glyco-, and lipo-proteins	Glycoproteins, glycolipids

number of residues per turn (N); N ≥10 for DNA, N ≤ 6 for proteins and N = 5–10 for glycans. Nucleotide strands run antiparallel whilst peptide and saccharide strands run either parallel or antiparallel. The tertiary structure describes the folded conformation of biomacromolecules. Its formation is inhibited by branch chains in glycans. Often the fold results in identifiable functional regions of the structures for nucleic acids and proteins (domains). The quaternary structure is rare in nucleic acids, though the association of rRNAs in an intact ribosome may be considered as an example. Proteins mediate the formation of quinternary structures between different biomacromolecules.

6.8 REFERENCES

Banin, E., Neuberger, Y., Altshuler, Y. *et al.* (2002) *Trends in Glycoscience Glycotechnology*, **14**, 127–37.

Bertozzi, C.R. and Kisseling, L. (2001) *Science*, **291**, 2357–64.

Bock, K. and Pedersen, C. (1983) *Advances in Carbohydrate Chemistry and Biochemistry*, **41**, 27–66.

Bohne-Lang, A., Lang, E., Foster, T. and von der Lieth, C.W. (2001) *Carbohydrate Research*, **336**, 1–11.

Boons, G.-J. (ed.) (1998) *Carbohydrate Chemistry*, Blackie, New York.

Bouveng, H. and Lindberg, B. (1960) *Advances in Carbohydrate Chemistry,* **15**, 58–68.

Caprioli, R.M., Malorni, A. and Sindona, G. (eds) (1996) *Mass Spectrometry in Biomolecular Sciences*, Kluwer Academic, Boston, MA.

Corfield, A.P. (ed.) (2000) *Glycoprotein Methods and Protocols: The Mucins*, Humana Press, NJ.

Davis, B.G. and Fairbanks, A.J. (2002) *Carbohydrate Chemistry*, Oxford University Press, Oxford, UK.

Dumitriu, S. (ed.) (2005) *Polysaccharides: Structural Diversity and Functional Versatility*, Marcel Decker, New York.

Dwek, R.A. (1996) *Chemistry Review*, **96**, 683–720.

Dwek, R.A., Edge, C.J., Harvey, D.J. *et al.* (1993) *Annual Reviews in Biochemistry*, **62**, 65–100.

Evans, J.N.S. (1995) *Biomolecular NMR spectroscopy*, Oxford University Press, London.

Fukuda, M. (1989) *Methods in Enzymology*, **179**, 17–29.

Fukuda, M.N., Sasaki, H., Lopez, L. and Fukuda, M. (1989) *Blood*, **73**, 84–9.

Geddes, R., Harvey, J.D. and Wills, P.R. (1977) *Biochemistry Journal*, **163**, 201–9.

Gidley, M.J. and Bociek, S.M. (1988) *Journal of the American Chemistry Society*, **110**, 3820–2.

Gorrin, P.A.J. (1981) *Advances in Carbohydrate Chemistry and Biochemistry*, **38**, 13–104.

Hakomori, S. (1964) *Journal of Biochemistry*, **55**, 205–208.

Helenius, A. and Aebi, M. (2001) *Science*, **291**, 2364–9.

Hess, K., Mahl, G. and Gütter, E. (1957) *Kolloid Z.* **155**, 1–9.

Hicks, K.S. (1988) *Advances in Carbohydrate Chemistry and Biochemistry*, **46**, 17–72.

Jiménez-Barbero, J. and Peters, T. (2003) *NMR Spectroscopy of Glycoconjugates*, Wiley-VCH, Weinheim, Germany.

Kaneko, K., Vey, M., Scott, M. (1997) *Proceedings of the National Academy of Sciences USA*, **94**, 2333–8.

Kobata, A. (1993) *Acc. Chemistry Research*, **26**, 319–24.

Marshall, J.J. (1974) *Adv. Carbohydrate Chem.* **30**, 257–370.

McGinnis, G.D. and Fang, P. (1980) *Methods in Carbohydrates*, **8**, 33–43.

McNaught, A.D. (1997) *Carbohydrate Research*, **297**, 1–90.

Millner, P. (ed.) (1999) *High Resolution Chromatography: A Practical Approach*, Oxford University Press, Oxford, UK.

Nosjean, O., Briolay, Y. and Roux, B. (1997) *Biochimera Biophysica Acta* **1331**, 153–86.

Patel, T.P. and Parekh, R.B. (1994) *Methods in Enzymology*, **230**, 57–66.

Rademacher, T.W., Parekh, R.B. and Dwek, R.A. (1988) *Annual Reviews in Biochemistry*, **57**, 785–838.

Reed, A.D. and Kerrett, R.J.S. (1968) *Carbohydrate Research*, **7**, 334–48.

Reinhold, V.N., Reihold, B.B. and Chan, S.M.E. (1996) *Methods in Enzymology*, **271**, 377–403.

Richter, W.J., Müller, D.R. and Domon, B. (1990) *Methods in Enzymology*, **193**, 609–23.

Roberts, G.C.K. (1993) *NMR of Macromolecules*, IRL Press, Oxford, UK.

Robyt, J.F. (1986) *Journal of Chemistry Education*, **63**, 560–1.

Romanowsk, A., Meunier, S.J., Tropper, F.D. *et al. (*1994) *Methods in Enzymology*, **242**, 90–101;

Rudall, K.M. (1963) A*dvances in Insect Physiology*, **1**, 257–313.

Sears, P. and Wong, C.-H. (2001) *Science*, 291, 2344–50.

Stack, R.J. and Sullivan, M.T. (1992) *Glycobiology,* **2**, 85–92.

Stowell, C.P. and Lee, Y.C. (1980) *Advances in Carbohydrate Chemistry and Biochemistry*, **37**, 225–81.

Szejtli, J., Augustat, S. and Richter, M. (1967) *Biopolymers*, **5**, 5–10.

Takasaki, S., Mizuochi, T. and Kobata, A. (1982) *Methods in Enzymology*, **83**, 263–8.

Tanretino, A.L. and Plummer, T.H. Jr. (1994) *Methods in Enzymology*, **230**, 44–57.

Tsuboi, S. and Fukuda, M. (2001) *BioEssays*, **23**, 46–53.

Varki, A., Cummings, R., Esko, J. *et al.* (1999) *Essentials of Glycobiology*, Cold Spring Harbor Laboratory Press, Cold Spring Habor.

Walkinshaw, M.D. and Arnott, S. (1981) *Journal of Molecular Biology*, **153**, 1055–62 and 1075–83.

World Wide Webs cited

CarbBank:	http://bssv01.lanes.ac.uk/gig/pages/gag/carbbank.htm
Consortium for Functional Glycomics:	http://www.functionalglycomics.org/static/consortium/
Carbohydrate structure suite (CSS):	http://www.dkfz.de/spec/css/
Glycosciences, tools:	http://www.glycosciences.de/tools/index.php

Glycosciences, sweet database:	http://www.glycosciences.de/sweetdb/index.php
IUPAC:	http://www.chem.qmw.ac.uk/iupac/
LINUCS:	http://www.dkfz.de/spec/linucs/ or http://glycosciences.de/tools/linucs/
NetOGlyc serve	http://www.cbs.dtu.dk/services/NetOGlyc/
SugaBase:	http://www.boc.chem.uu.nl/sugabase/database.html

CHAPTER 7

STUDIES OF BIOMACROMOLECULAR STRUCTURES: SPECTROSCOPIC ANALYSIS OF CONFORMATION

7.1 BIOCHEMICAL SPECTROSCOPY: OVERVIEW

Spectroscopic methods are used in the structural characterization of biomolecules (Bell, 1981; Campbell and Dwek, 1984; Greve *et al.*, 1999; Hammes, 2005). These methods are usually rapid and noninvasive, require small amount of samples, and can be adapted for analytical purposes. Spectroscopy is defined as the study of the interaction of electromagnetic radiation with matter, excluding chemical effects (photochemistry refers to the interaction with chemical effects). The electromagnetic spectrum covers a wide range of wavelengths (Figure 7.1).

The interaction of electromagnetic radiation with matter gives rise to scattering, absorption or emission of the radiation. Scattering is usually detected by measuring the intensity of radiation at some angle θ to the incident wave (turbidity refers to measuring the reduced transmitted light at $\theta = 0$). Electrons are the usual scatterers in molecules, while nuclei scatter neutrons. If the scattered radiation has the same frequency as the incident radiation, the scattering is said to be elastic (conservation of energy), otherwise the scattering is inelastic (changes in frequency). Three cases will be considered:

1. *Refraction and reflection*: Refraction results when light is scattered in the same direction as that of the incident light. The wavelength (λ) of the incident radiation is much greater than the dimension of large arrays of essentially rigid particles, e.g. crystals, very little scattering is observed other than $\theta = 0$. Reflection then results when light is scattered in the direction opposite to that of the incident light.
2. *Diffraction*: When the wavelength of the incident radiation is much less than the dimension of arrays in the crystals, three-dimensional interference patterns usually called diffraction patterns are generated. The diffraction pattern gives information about the lattice and the constituent molecules of the array.
3. *Solutions*: When the dimension of a particle in solution is much less than λ, scattering observed at $\theta = 0$ is related to the concentration and size of the particle. When the dimension of the scatter is greater than λ, angular dependent scattering can be measured to provide information about the size and shape of the particle.

Absorption is usually measured by varying the frequency (or wavelength) of the applied radiation. The frequency dependence of absorption arises because energy is

---

Biomacromolecules, by C. Stan Tsai
Copyright © 2007 John Wiley & Sons, Inc.

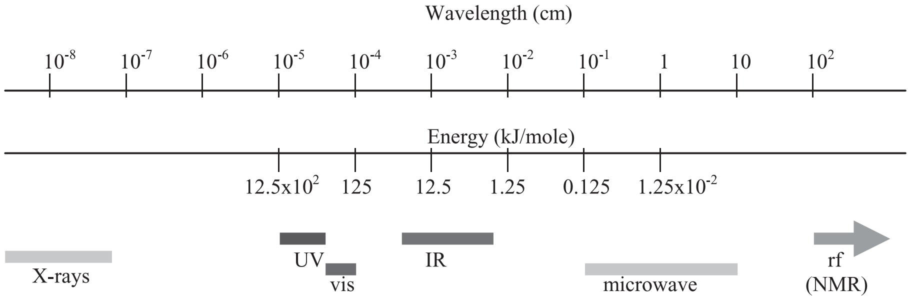


Figure 7.1 Electromagnetic radiation and its corresponding spectra

absorbed by transition induced between different energy states of the molecules in the sample. Each molecule is associated with many types of energies of which the most important are:

$$E_{total} = E_{electronic} + E_{vibration} + E_{rotation} + E_{translation} + E_{electron\ spin\ orientation} + E_{nuclear\ spin\ orientation}$$

Each of these energies is quantized into energy levels that are characteristic states of the molecule. The ground state is defined as the state of the lowest energy and states of higher energy are called excited states, that is, they are said to be degenerate if two or more states of the molecule have the same energy. A molecule will absorb radiation only when the frequency (ν) of the radiation is related to the energy difference (ΔE) between two energy levels by the equation:

$$\Delta E = h\nu,$$

where h is Planck's constant (6.67×10^{-27} erg s). The frequency is related to the wavelength by $\nu = c/\lambda$, in which c is velocity of light in vacuum (3×10^{10} cm s^{-1}). Another factor (selection rule) for predicting the transition probability is that there must be a charge displacement between one energy state to another. Only those components of the electromagnetic radiation that are in the same direction as the transition dipole moment will cause transitions. If the angle between the direction of the applied wave and the direction of the transition moment is θ, then the effective value of the transition moment is proportional to $\cos\theta$ and the transition probability is proportional to $\cos^2\theta$. Two types of charge displacements during the transition are electric transition dipole (μ_e) and magnetic transition dipole (μ_m). The applied electromagnetic radiation interacts with these dipoles to cause changes in energy states; the electric component interacts with μ_e and the magnetic component interacts with μ_m, and the larger the transition dipole moment, the larger will be the transition probability.

The selection rules help to predict the probability of a transition but are not always strictly followed. If the transition obeys the rules it is allowed, otherwise it is forbidden. A molecule can become excited in a variety of ways, corresponding to absorption in different regions of the spectrum. Thus certain properties of the radiation that emerges from the sample are measured. The fraction of the incident radiation absorbed or dissipated by the sample is measured in optical (ultraviolet and visible) absorption spectroscopy and some modes of nuclear magnetic resonance spectrometry (NMR). Because the relative positions of the energy levels depend characteristically on the molecular structure, absorption spectra provide subtle tools for structural investigation.

Emission of radiation is measured at some angle θ to the incident beam as a molecule changes from an excited energy state to a lower energy state. A molecule can change its energy from a higher excited state to lower one by three processes:

1. *Simulated emission*: A light amplification by stimulated emission of radiation (LASER) operates in systems where a nonequilibrium distribution of energies is created by a pump that induces transition to a higher excited state. As a result, the emission of some radiation is made to stimulate a cascade of emission. This emission will stop when the equilibrium to the population of energy states is returned.
2. *Thermal* (*radiationless*) *emission*: The common way for a molecule to return to a lower energy state is by the liberation of heat via collision, vibration and molecular motion in the intermolecular and intramolecular de-excitation processes.
3. *Spontaneous emission*: The molecule acts an oscillator and radiates its energy hν without any other interaction with its environment.

The measurement of the emitted radiation at a wavelength other than that used for the excitation form the basis for fluorescence, phosphorescence and Raman scattering spectroscopies.

In addition to the intensity, other properties such as polarization are concerned in optical rotatory dispersion and circular dichroism. The various processes give rise to different spectroscopic methods, as summarized in the Table 7.1. Various spectra (UV, IR, NMR and MS) of simple biomolecules can be accessed from Spectral Database Systems (SDBS) of the National Institute of Material and Chemical Research, Japan at http://www.sist.go.jp/RIODB/SDBS/menu-e.html.

Several factors must be considered for a particular biomacromolecular structure application that will affect the choice of spectroscopic methods. These include structural resolution necessary, chemical nature of biomacromolecule (protein, nucleic acid, or glycan), amount/concentration of biopolymer available, sample preparation (solid or solution), solvents of interest, and desired structure information (secondary or tertiary structure). Structural resolution varies considerably for the various spectroscopic methods, with X-ray diffraction and NMR providing atomic resolution (high resolution) and ultraviolet (UV) absorption revealing merely information about the polarity of the chromophore's environment (low resolution). X-ray studies require crystals while NMR experiments prefer solutions in deuterated solvent. Solvent preferences can affect the choice of spectroscopic method as, for example, infrared (IR) encounters strong interference from water, while optical rotatory dispersion (ORD) and circular dichroism (CD) do not. Some of the commonly used spectroscopic methods in structural analyses of biomacromolecules will be discussed.

7.2 ULTRAVIOLET AND VISIBLE ABSORPTION SPECTROSCOPY

7.2.1 Basic principles

Ultraviolet (UV) and visible spectra. also known as electronic spectra. involve transitions between different electronic states. The electronic transition is accompanied by the vibrational and rotational transitions so that what would otherwise be an absorption line becomes a broad peak containing vibrational and rotational fine structure. Furthermore, the molecular interaction between solute and solvent levels it to a smooth curve (envelope) for the absorption spectra in solutions. The accessible regions are 200–400 nm for

TABLE 7.1 Various spectroscopic methods used in biomacromolecular characterization

Spectroscopy	Principle	Biopolymer	Structural information and application
Ultraviolet and visible	Absorption of UV and visible radiation leading to electronic excitation	N, P	Ligand binding sites, side chain exposure, environment. Applicable to proteins and nucleic acids but not glycans. Quantitative analysis. DNA conformation. Annealing and hybridization studies of nucleic acids.
Fluorescence	mission of radiation when a molecule in an excited electronic state returns to the ground state	N, P	Environment, relative abundance and interactions of fluorophore. Quantitation of proteins and fluorophoric compounds. Ligand binding studies.
Infrared	Absorption of IR radiation leading to vibrational excitation	N, P, G	Presence and environment of functional groups. Structural diagnosis such as conformation, inter- and intra-molecular/chain hydrogen bonding. Identification of unknown by fingerprinting. Studies of biomacromolecular dynamics by H-D exchange.
Raman	Scattering of visible or UV light with abstraction of some of the energy leading to vibrational excitation	N, P, G	Similar to infrared but with sampling restriction. However, aqueous solution can be used. Functional groups which have weak IR absorption often give strong Raman spectra
Nuclear magnetic Resonance	Absorption of radiation giving rise to transitions between different spin orientations of nuclei in a magnetic field	N, P, G	Sequences of biomacromolecules. Complete 3D structures. Unambiguous detection of certain functional groups and information about the environment. Structural identification by fingerprinting. Studies of biomacromolecular interactions with ligands.
Electron spin Resonance	Absorption of radiation giving rise to transitions between opposite spin orientations of unpaired electrons in a magnetic field	P	Detection and estimation of free radicals and diagnosis of their structure and electron distribution of metalloproteins. For other biomacromolecules, probes are needed.
Optical rotatory Dispersion	Rotation of the plane polarized light by asymmetric molecules in solution without and with variation in wavelength	P, G	Determination of relative and absolute configurations of asymmetric centers. Location of functional groups in certain types of compounds. Information about conformation.
Circular dichroism	Difference in intensity of absorption of right- and left-circularly polarized light by functional groups in asymmetric environment	N, P, G	Applications similar to but more powerful than ORD especially for functional groups such as C=O. Conformational analysis of biomacromolecules in solutions.
X-Ray Diffraction	Interference between scattered X-rays caused by atomic electrons	N, P, G	Determinations of complete 3D structure and stereochemistry. Such structural analysis gives bond lengths, and angles as well as distances between non-bonded atoms in crystals.
Neutron diffraction	Interference between scattered neutron beam caused by atomic nuclei	N, P, G	Particularly useful for the location of hydrogen atoms in a molecule and insoluble macromolecules.

Note: Biopolymers refer to the majority of native biomaromolecules of that class, i.e. N (nucleic acids), P (proteins) and G (glycans).

UV and 400–750 nm for visible spectra. The groups giving rise to the electronic transitions in the accessible regions are termed chromophores, which include aromatic amino acid residues in proteins, nucleic acids and derivatives, nucleotide coenzymes (e.g. NAD(P)H), flavins, hemes and some transition metal ions. Polysaccharides are devoid of chromophores and are therefore UV and visible spectrally inactive.

Spectral lines are not infinitely sharp (i.e. truly monochromatic). Various factors, such as the lifetime of the excited state and presence of the overlapping bands, contribute to the broadening of a spectral line. In general, short-lived energy states give broad spectral lines, while long-lived states give narrower spectral lines.

Two parameters characterize an absorption band, namely the position of peak absorption (λ_{max}) and the extinction coefficient (ε), which is related to concentrations of the sample by the Beer–Lambert law:

$$A = \log(I_0/I_s) = \varepsilon c l$$

where I_0 and I_s are the incident and transmitted radiation respectively and l is the length of the cell through which radiation travels.

The electronic energy levels of molecules are described by molecular orbitals (MOs). In an electronic transition, an electron is transferred from one MO to another. For electronic spectroscopy of biomacromolecules, three MOs associated with unsaturated centers are important. These are the π bonding orbital, the π^* antibonding orbital, and the n nonbonding (lone pair) orbital. Transitions between electronic energy levels are allowed if they result in an unsymmetrical movement of charge, such as $\pi \rightarrow \pi^*$ transitions. The orbitals in a molecule can be distorted by mixing or contamination with other orbitals. This can make forbidden transitions allowed, such as the simultaneous transitions involving both electronic and vibrational excitation (known as vibronic). The $n \rightarrow \pi^*$ transition of $C{=}O$ (~290 nm) is an example of forbidden transition that becomes allowed. One of the constraints in the absorption spectroscopy of biomacromolecules is the need to work under/near physiological condition (in a solvent buffered at a pH near 7.0 and containing sufficient electrolyte). The use of aqueous solvent restricts absorption spectral measurements to wavelengths longer than 170 nm. Changes in the environment of the chromophore may lead to shifts in absorption maximum (λ_{max}) and absorption intensity (ε) of an electronic spectrum of a macromolecule. The absorption maximum may shift toward longer wavelength (red shift or bathochromic shift) or shorter wavelength (blue shift or hypochromic shift). The shift may lead to a decrease in the absorption intensity (hypochromic effect) or an increase in the absorption intensity (hyperchromic effect). A point common to all curves produced in the spectra of a compound taken at several experimental conditions (e.g. changes in pH, temperatures or solvents) is known as an isosbestic point. The electronic transitions for the chromophores of proteins and nucleic acids will be considered.

7.2.2 Amino acid residues and peptide bonds

Three classes of chromophores contribute to protein absorption; namely the peptide bond, amino acid side chains and any prosthetic groups. The $n \rightarrow \pi^*$ and $\pi \rightarrow \pi^*$ transitions of the backbone peptide group occur in the far-UV range. The $n \rightarrow \pi^*$ forbidden transition occurs at 190–220 nm ($\varepsilon_{max} = 100$), while the main $\pi \rightarrow \pi^*$ is observed at ~190 nm ($\varepsilon_{max} = 7000$). α-Helix has the lowest absorption intensity among the three secondary structures; the hyperchromicity of proteins in the peptide absorption region (190–210 nm) is taken as a lowering helical content. A number of amino acid side chains with unsaturation centers (e.g. Asp, Asn, Glu, Gln, His, Arg) have transitions at ~210 nm but are not observed

TABLE 7.2 Spectral parameters for selected amino acids at pH 7.0.

Amino acid	λ_{max} (nm)	ε_{max} (cm^2·mol) × 10^{-3}
Cystine	250	0.3
Histidine	211	5.9
Phenylalanine	257	0.2
	206	9.3
	188	60.0
Tyrosine	274	1.4
	222	8.0
	193	48.0
Tryptophan	280	5.6
	219	47.0
Proteins: α-Helix	189–204	4.1
β-Sheet	187	7.7
Coil	190	6.9

because of the intensely absorbing peptide backbone groups, which have significant absorption up to ~230 nm. The most useful UV range for proteins is at wavelengths greater than 230 nm where the aromatic side chains of Phe, Tyr and Trp contribute to the absorption (Table 7.2).

The contribution of Phe in the range is minimal if Tyr and Trp are present in a protein. Trp has the most intense absorbency in this range, which is the basis for the measurement of protein concentrations by UV spectrometry. The number of moles of Tyr and Trp per mole of protein, M_{tyr} and M_{trp} can be determined from the absorption of the protein solution (6 M guanidine hydrochloride in 0.02 M phosphate buffer, pH 6.5) measured at 288 and 280 nm:

$$\varepsilon_{288} = 4815M_{trp} + 385M_{tyr} \quad \text{and} \quad \varepsilon_{280} = 5690M_{trp} + 1280M_{tyr}.$$

Variations in pH affect the spectra of Tyr and Trp, especially the deprotonated Tyr whose shift can be monitored at 295 nm. Another chromophore is the disulfide group of cystine, which has a very week absorption at 250 nm. It has been common practice to determine protein concentrations spectrometrically at 280 nm by assuming an absorbency of 1.0 for a 1 mg/mL protein solution.

Some of the prosthetic groups displaying electronic absorption in the region of 400–600 nm such as flavins ($\varepsilon_{455} \approx 1.27 \times 10^4$), pyridoxal phosphate ($\varepsilon_{415} \approx 0.65 \times 10^4$) and hemes ($\varepsilon_{550} \approx 2.77 \times 10^4$). Their absorption bands are usually sensitive to the local environment and can be used to monitor structural and environmental changes.

7.2.3 Purines, pyrimidines and nucleic acids

The absorption of nucleic acids also arises from $n \rightarrow \pi^*$ and $\pi \rightarrow \pi^*$ transitions of purine and pyrimidine bases. The spectra of nucleic acids and derivatives occur around 200–300 nm (Table 7.3). The spectra of all four nucleosides are sensitive to pH. Protonation of C and G results in a large red shift. Deprotonation of U and T also results in a large red shift. In a typical nucleic acid, the spectral properties of the isolated chromophores merge into a smooth single band with $\lambda_{max} \approx 260$ nm. The average molar extinction coefficient of a nucleoside at 260 nm is about 1×10^4 cm^2·mol. Hence the UV spectrum of a nucleic acid can be measured at concentrations fairly accurately as low as 3 μg/mL.

TABLE 7.3 Spectral parameters for nucleosides and nucleic acids

Nucleoside and nucleic acid	λ_{max} (nm)	ε_{max} (cm^2·mol) × 10^{-3}
Adenosine	259.5	14.9
Guanosine	276	9.0
Cytidine	271	9.1
Uridine	261	10.1
Thymidine	267	9.7
DNA	258	6.6
RNA	258	7.4

The absorption intensity of nucleic acids is lower than in mixtures of their component monomers. The hypochromicity exhibited by nucleic acids arises from the interaction between electronic states of nitrogen bases due to their constrained polymeric structures, i.e. $\varepsilon_p < \Sigma x_i \varepsilon_{mi}$, where ε_p is the extinction coefficient of nucleic acid, x_i and ε_{mi} are the mole fraction and extinction coefficient of the ith monomer component. The percent hypochromicity at λ (H_λ) is defined as

$$H_\lambda = 100(1 - \varepsilon_{p,\lambda}/\Sigma x_i \varepsilon_{mi,\lambda})\%$$

For polynucleotides, an hypochromic effect of the order of 10–50% is frequently observed. The intact double helix DNA absorbs roughly 30% less than a mixture of the component monomers. A few of minor bases found in tRNA show $\lambda_{max} > 300$ nm, such as the Y base ($\lambda_{max} = 325$ nm) and 4-thiouridine ($\lambda_{max} = 340$ nm), which can be exploited to investigate these molecules.

7.2.4 Perturbation difference absorption spectroscopy

The biochemical applications of UV and visible spectroscopy are determination of concentrations, interactions of ligands with biomacromolecules and conformational changes caused by experimental perturbations. The sensitivity of UV and visible spectra to the solvent environment of the chromophore leads to shifts in the absorption maximum and the absorption intensity, and it is the basis of solvent perturbation spectra in the structural studies of biomacromolecules (Donovan, 1969). The idealized development of environmental (solvent) contribution to the extinction coefficient of a chromophore is described by

$$\varepsilon_s = \varepsilon_0[1/n\{(n^2 + 2)/3\}^2]$$

where ε_s, ε_0 are extinction coefficients in the presence and absence of solvent, and n is the refractive index of the solvent. This accounts for the increase in extinction usually observed with chromophores as the refractive index is increased. Shifts in absorption spectra accompanied conformational changes are generally monitored by the perturbation difference absorption spectroscopy.

Spectroscopic observation of perturbations of the chromophores of biomacromolecules has become a valuable method for the determination of the conformation of native and altered structures in solution. When perturbation of the chromophores of a biopolymer is observed, the resulting conformational changes may be detected by measurement with any one of the spectroscopic methods, of which absorption spectroscopy will be considered. To record perturbation of difference spectrum, unperturbed preparation is used as the reference and perturbed preparation as the sample. Perturbation of a chromophore results in a shift of its absorption bands along the wavelength axis, accompanied by an increase or decrease in integrated absorption. To a first approximation, neglecting absorption changes compared to the wavelength shift, the difference spectrum is represented by

$$\Delta\varepsilon(\lambda) \approx -\Delta\lambda(d\varepsilon/d\lambda) \quad \text{if } \Delta\lambda \text{ is small.}$$

The relationship describes the magnitude of the difference extinction ($\Delta\varepsilon$) and the magnitude of the wavelength shift producing it ($\Delta\lambda$), in terms of the slope of the absorption curve of the chromophore ($d\varepsilon/d\lambda$). All the difference spectra resemble the first derivatives of the absorption spectra of the respective chromophores. However, an increase in absorption accompanying a red shift in the spectrum often smear the negative portion of the difference spectrum.

7.3 FLUORESCENCE SPECTROSCOPY

Fluorescence is the emission of radiation that occurs when an excited molecule returns to the ground state (Lakowicz, 1999). The relaxation of the excited molecules back to the ground state may occur by the radiationless process. Its presence results in a quenching of the fluorescence intensity. Fluorescence involves two processes: absorption and subsequent emission. The shape of the emission band is approximately a mirror image of the absorption band if the vibronic structures of the two states are similar. Each process occurs in the time scale given by the inverse of the transition frequency ($\sim 10^{-15}$ s), but there is a time lag of about 10^{-9} s when the molecule exists in the excited state. Fluorescence occurs at a lower frequency than that of the exciting light. Because the detection frequency is different from the incident frequency, the sensitivity in fluorescence spectroscopy is high (sample concentration in the 10^{-8} M range). Therefore fluorescence spectroscopy is widely used in the analysis of fluorescence-tagged biomolecules. It is the detection method of choice for automatic sequence analyses and microarray analyses of nucleic acids and proteins.

The dependence of the fluorescence intensity on the wavelength of the exciting light is known as the excitation spectrum, while the variation of the fluorescence intensity with the wavelength of the emitted light is referred to as the emission spectrum.

The molecular group giving rise to fluorescence is termed fluorophore (fluorescence chromophore). The main fluorophores can be classified into natural fluorophores such as tryptophan residue in proteins, NAD(P)H, FMN/FAD and fluorescent indicators (probes) such as dansyl chloride, 8-anilino-1-naphthalene sulfonate (ANS), and ethidium bromide. Nucleic acids do not have appreciable fluorescence, except for a minor base (Y-base) in tRNA (Table 7.4).

The measurable parameters are the quantum yield (ϕ_F), and the intensity and the position of peak emission (λ_{max}). The quantum yield or fluorescence efficiency is the fraction of molecules that becomes de-excited by fluorescence and is defined as

$$\phi_F = \tau/\tau_F$$

where τ is the observed lifetime of the excited state and τ_F is the radiative lifetime that can be related to the molar extinction coefficient (ε_{max}) by $1/\tau_F \approx 10^4\varepsilon_{max}$. The quantum yield is sensitive to the immediate surroundings of the fluorophore and to specific quenching processes. The measured fluorescent intensity (F_λ) depends on the initial population of the excited state (I_A) and the quantum yield by

$$F_\lambda = I_A\phi_F$$

TABLE 7.4 Characteristics of some natural fluorophores

Natural fluorophore	Trp Aqueous, pH 7	Tyr Aqueous, pH 7	Phe Aqueous, pH 7	Y-base Yeast $tRNA^{phe}$
Absorption:				
λ_{max} (nm)	280	274	257	320
ε_{max} ($\times 10^{-3}$)	5.6	1.4	0.2	1.3
Fluorescence:				
λ_{max} (nm)	348	303	282	460
ϕ_F	0.2	0.1	0.04	0.07
τ_F (ns)	2.6	3.6	6.4	6.3
Sensitivity:				
$\varepsilon_{max}\phi_F$ ($\times 10^{-2}$)	11	1.4	0.08	0.91

Applications of fluorescence spectroscopy include ligand binding, probing of environment and measurement of distance between fluorophores. Fluorescence is very sensitive to the environment and the various parameters (e.g. λ_{max}, ϕ_F and τ) that are affected, can be exploited for structural studies:

- Effect on emission maximum (λ_{max}): The effect can be used to estimate polarity of the environment around the fluorophore. In general, the excited state is more polar than the ground state. Thus the excited molecules tend to interact with a polar environment favorably and cause the red shift in the emission spectrum. However, an orientation constraint on the fluorophore may cause a blue shift if the probe molecules do not have time to undergo rearrangement.
- Effect on quantum yield (ϕ_F): In general, the quantum yield (and measured fluorescence intensity) of a fluorophore increases as the polarity of its environment decreases, due possibly to a reduction in the rate of intersystem crossing into a nonpolar environment.

Empirical rules for interpreting fluorescent spectra of proteins:

- All fluorescence of a protein is due to Trp, Tyr and Phe, unless the protein is known to contain fluorophoric prosthetic group.
- The λ_{max} of the Trp fluorescence spectrum shifts to shorter wavelengths (blue shift) and the intensity of λ_{max} increases as the polarity of the environment decreases. Otherwise Trp is internal in a nonpolar environment or the solvent induces conformational change that brings it to the surface.
- If Trp or Tyr residues are in a polar environment, their ϕ_F decreases with increasing temperature, whereas in a nonpolar environment, little change accompanies the temperature fluctuations.
- If a substance binds to a protein and Trp fluorescence is quenched, either the binding induces a gross conformational change or the presence of Trp in/near the binding site.
- If a substance known to be a quencher, e.g. I^-, NO_3^- or Cs^+ quenches Trp or Tyr fluorescence, the amino acid residues must be on the surface of the protein. Otherwise these residues may be internal, in a crevice or a highly charged region of the protein.
- If a substance that does not affect the quantum yield of the free amino acid affects the fluorescence of a protein, it must do so by producing conformational change in the protein.

There are two types of quenching commonly encountered. The first type derives from collision processes and this collision quenching has a short-range effect. It depends on the rate of diffusion of the fluorophore molecules and is controlled by factors such as molecular size and media viscosity. The second type of quenching arises form nonradiative process called resonance energy transfer from one chromophore to another. It has a long-range effect (up to a distance of ~5 nm between the chromophores) without emission and reabsorption of radiation. The most common type of energy transfer is from the excited singlet state of a donor to the excited singlet state of an acceptor. The energy separation in each case must match (in resonance) as indicated by the overlap of the fluorescence spectrum of the donor and the absorption spectrum of the acceptor. Analysis of this quenching allows the measurement of the distance between the chromophores.

In the nonradiative quenching, the efficiency (E_T) of depopulation by resonance energy transfer is

$$E_T = k_T/(1/\tau + k_T) = 1 - \tau_T/\tau = 1 - \phi_T/\phi_D$$

where τ and τ_T are lifetimes (reciprocal to the rate constants, k and k_T) of depopulation in the absence and presence of resonance energy transfer. ϕ_D and ϕ_T represent the quantum yields of the donor in the absence and presence of the energy transfer respectively, since the rate of resonance energy transfer (k_T) is proportional to the intermolecular distance by $1/R^6$. Thus

$$E_T = R^{-6}/(R_0^{-6} + R^{-6}) = R_0^6/(R^6 + R_0^6)$$

where R_0 is the distance at which the energy transfer is 50% efficient (i.e. $1/\tau = k_T$). It becomes

$$R = R_0\{(1 - E_T)/E_T\}^{1/6} = R_0\{\phi_T/(\phi_D + \phi_T)\}^{1/6}$$

Therefore a measurement of the quenching of the fluorescence (ϕ_T/ϕ_D) in the presence and absence of the acceptor, R can be calculated if R_0, which is constant for each donor-acceptor pair, is known.

The major purine and pyrimidine bases in nucleic acids have only a very low fluorescence of practical value. However, fluorescence methods can be applied to studies of nucleic acids by substituting a fluorescent analog for a normal base or by binding a fluorophore. For example, fluorophore 2-aminopurine can be substituted for adenine. This adenine analog hydrogen bonds to thymine without distorting the normal B-DNA structure. Proflavine monosemcarbazide can be covalently attached to the 3′-end of RNA. A fluorescence probe, ethidium bromide is intercalated into DNA and the fluorophore is excited with UV light to emit visible light. Ethidium fluorescence is strongly quenched in water, but not when shielded in the hydrophobic environment of the DNA helix.

Fluorescence detection employs direct detection methodology that is simple, sensitive and easy to automate. Most of the dyes (fluorophores) used for biomacromolecular analysis have Stokes shift values (differences between the absorption maximum and the emission maximum) in the range of 20–60 nm. The greater the distance between the absorption and emission spectra, the easier it is to separate the fluorescent signal from the excitation light. For this reason, dyes with large Stokes shifts are valuable for biochemical applications such as sequence and microarray analyses (subsections 4.2.3 and 14.3.4).

7.4 INFRARED SPECTROSCOPY

7.4.1 Basic principles

The energy of most molecular vibrations corresponds to that of the IR region of the electromagnetic spectrum. Molecular vibrations may be detected and measured either in an IR spectrum or indirectly in a Raman spectrum. The common region of the IR spectrum is 2.5–16 μ (4000–625 cm^{-1}), which most IR spectrometers cover. Infrared light is absorbed when the oscillating dipole moment (due to a molecular vibration) interacts with the oscillating electric vector of the infrared beam. A simple rule for deciding if this interaction (and hence absorption of light) occurs is that the dipole moment at one extreme of a vibration must be different from the dipole moment at the other extreme of the vibration. Thus the selection rule for a vibration to be IR active is that the vibration must result in a change in the dipole moment. The important consequence of this selection rule is that in a molecule with a center of symmetry, those vibrations symmetrical about the center of symmetry are inactive in IR and active in Raman. Those vibrations that are not centrosymmetric are inactive in Raman and usually active in IR. Transitions from the ground vibrational state to the first excited vibrational state are called fundamentals (or first harmonics). Since only fundamentals are strictly allowed, these constitute the dominant transitions. Those transitions to the higher levels are called overtones (second harmonics), which are usually not observed except when hydrogen bonding is present. Associated with each vibrational energy level are many closely spaced rotational energy levels, which are unresolved in solution broadening the vibrational band (Parker, 1983; Mantsch and Chapman, 1996).

A complex molecule has a large number of vibrational modes. The number of modes of normal vibration is 3N – 6 for a molecule consisting of N atoms. For biomacromolecules there are a great number of vibrational transitions. However, many of the vibrations can be localized to particular bonds or groups (e.g. O—H, N—H and C═O groups) because they involve a large displacement of just two bonded atoms with little interaction from any other vibrations. These result in the observed characteristic group frequencies (Table 7.5), which are the basis of most applications of IR spectroscopy.

Water has a high absorbency throughout the IR region and thus interferes with IR spectra of biomolecules, which are usually recorded in aqueous solutions. For aqueous solutions, the water-insoluble cells (made of CaF_2 or LiF) and double-beam instrument. which allows compensation for the water absorbency (with water in the reference cell), are used.

The main experimental parameter in IR is the frequencies (e.g. ν for stretching, δ for in-plane bending and γ for out-plane bending) of the absorption bands characterizing functional groups, in particular $\nu_{C=O}$ and ν_{N-H} for biomolecules. The vibrational frequency ν_{vib} of a bond can be approximately described by the formula

$$\nu_{vib} = 1/(2\pi)[k/\mu]^{1/2}$$

where k is the force constant and μ is the reduced mass, i.e. $\mu = m_1m_2/(m_1 + m_2)$ for the two bonded atoms with masses m_1 and m_2.

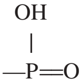
The shape of the IR absorption band is determined primarily by collision broadening and will approximate to a Gauchy function:

$$A(\nu) = A_m/\{1 + [2(\nu - \nu_m)/\Delta\nu_{1/2}]^2\}$$

in which the half-bandwidth ($\Delta\nu_{1/2}$) is usually in the range of 2–20 cm^{-1} in liquid systems.

With optically anisotropic samples (crystals and oriented films and fibers), dichroism measurements indicate the direction of the transition dipole moment. Since the

TABLE 7.5 Characteristic group IR frequencies of biochemical interest

	Group	IR band (cm^{-1})	Remark
Hydroxy	Water (solution)	3710	Liquid/solution
	Water (crystal)	3600–3100	Solid state spectra
	Free —OH	3650–3200	$\nu_{\text{O—H}}$, sharp
		1410–1260	$\gamma_{\text{O—H}}$.
	Hydrogen bonded —OH	3600–3200	$\nu_{\text{O—H}}$, often broad but may be sharp for some intramolecular H-bonds. The lower the frequency the stronger the H-bond.
Amine/imine	—NH_2, >NH, =NH	3500–3300	$\nu_{\text{N—H}}$. Primary amines show two bands in this range while secondary amines absorb weakly. The imine =N—H band is sharp.
		1650–1940	$\gamma_{\text{N—H}}$, medium for primary amines and weak for secondary amines
	—NH_3^+ (amino acid)	3130–3030	$\nu_{\text{N—H}}$, sometimes accompanied by broad band near 2500 and 2000 cm^{-1}.
	=NH^+—	2700–2250	$\nu_{\text{N—H}}$, broad
Amide	—$CONH_2$	3500, 3400	$\nu_{\text{N—H}}$, lowered by ~150 cm^{-1} on H-bonding.
	—CONH—	3460–3400 3100–3070	$\nu_{\text{N—H}}$. Two bands and lowered on H-bonding and solid state samples
Thiol	—SH	2600–2550	$\nu_{\text{S—H}}$, weaker than $\nu_{\text{O—H}}$ and less affected by H-bonding.
Aldehyde	—CHO	1740–1720	$\nu_{\text{C=O}}$, strong. Shift to 1655–1625 cm^{-1} due to intramolecular H-bonding.
Ketone	>CO	1725–1705	$\nu_{\text{C=O}}$, strong. Lowered by 10–20 cm^{-1} due to H-bonding and in solid state.
	Amino/hydroxy>CO	1655–635	$\nu_{\text{C=O}}$, low due to intramolecular H-bonded to amino/hydroxy.
Carboxyl	Acid, —COOH	3000–2500	$\nu_{\text{O—H}}$, lower by H-bonding.
	Aliphatic acid, —COOH	1725–1700	$\nu_{\text{C=O}}$, strong.
	Aromatic acid, —COOH	1700–1680	$\nu_{\text{C=O}}$, strong.
	Ion, —COO^-	1610–1550	$\nu_{\text{C=O}}$, strong.
	Amide, —CO—N<	~1690	$\nu_{\text{C=O}}$, strong. Lowered to ~1650 cm^{-1} due to H-bonding or in solid state.
	Ester, —CO—OR	1750–1735	$\nu_{\text{C=O}}$, strong. α-keto esters occur at 5 cm^{-1} higher while H-bonded β-keto esters lower to ~1650 cm^{-1}.
Phosphate	 $\overset{\text{OH}}{\overset{\vert}{\text{—P=O}}}$	1240–1180	$\nu_{\text{P=O}}$, strong
		2700–2560	Hydrogen bonded OH

Note: stretching and bending frequencies are represented by ν and γ respectively.

absorption of IR (the electromagnetic radiation) by a molecule depends on the angle, θ between the direction of the applied radiation and the direction of the transition moment, IR measurements can be made using plane-polarized light in two different directions to give two absorption values, e.g. $A_{\|}$ and $A_{\perp}$. In solution, $A_{\|}$ and $A_{\perp}$ are equal because a molecule takes up all possible orientations in its transition dipole moment. But in a system of oriented molecules, $A_{\|}$ and $A_{\perp}$ may be different, known as dichroism (linear dichroism for the plane-polarized light), which gives information on the direction of the transition dipole moment correlating to the molecular conformation.

If the polymeric molecules are perfectly aligned with the fiber axis, the ratio (R_0) of the integrated intensities measured with the plane-polarized light (electric vector vibrating parallel and perpendicular to the fiber axis) and the inclination (α) of the transition moment direction to the fiber axis is related by

$$R_0 = 2\cot^2\alpha.$$

In instances where the transition moment direction is either parallel ($\alpha = 0°$) or perpendicular ($\alpha = 90°$) to the chain (fiber) axis, the corresponding values will be $R_0 = \infty$ and 0 respectively. An important practical aspect of interpreting dichroism is that in linear polymers, which possess a regularly repeating structure, the localized vibrations of the monomer units are coupled and each mode of the monomer gives rise to three normal modes of vibration of the molecule. With helical polymers, the transition moment directions associated with one of these modes will be parallel to the molecular axis and other two will be perpendicular to that axis. The half-bandwidths of the three bands can be markedly different.

7.4.2 Biochemical applications

Analysis of the characteristic group frequencies forms the basis of most applications of IR spectroscopy in biochemical systems. The main applications involve monitoring vibrations of selected groups, either on ligands or biomacromolecules in the ligand bindings, probing hydrogen bonds and molecular conformation in structure perturbation experiments (Singh, 2000).

Methods currently being used to extract information on protein secondary structure from IR spectra are based on empirical correlation between the frequencies of certain vibrational modes and types of secondary structure of polypeptide chains. Infrared spectra of proteins are dominated by absorption bands associated with vibrations of the main chain, and the strong bands around 1550, 1650 and 3300 cm^{-1} identified as N—H bending, C═O stretching and N—H stretching vibrations respectively. The vibrational modes associated with these main chain bands are conformation sensitive (Timasheff and Gorbunoff, 1967). These bands are fairly easy to measure in D_2O solutions because each peptide unit contributes and D_2O does not absorb strongly in these regions.

The mode most often used and best characterized is the so-called amide I band. It represents primarily the C═O stretching vibrations of the amide groups (coupled to in-plane bending of the N—H and stretching of the C—N bonds) and gives rise to IR band(s) in the region between 1600 and 1700 cm^{-1}. The amide II band is polarized nearly parallel to the C—N peptide bond, and thus nearly perpendicular to the N—H bond. Hydrogen bonding shifts the energies of the three peptide vibrations (Table 7.6). The two stretching bands are moved to lower energy because the presence of a hydrogen bond makes it easier to stretch the carbonyl oxygen toward the hydrogen-bonding donor, or the amide nitrogen toward the acceptor. The bending band (amide II) is shifted to a higher energy by hydrogen bonding because hydrogen bonds are roughly linear, which resists the bending.

TABLE 7.6 Characteristics of main chain IR bands of proteins

Vibration mode (cm^{-1})	Oscillating transition dipole	Hydrogen-bonded (cm^{-1})				Non-H-bonded (cm^{-1})
		α Helix	Dichroism	β Sheet	Dichroism	
N—H Stretching	←N—H→	3290–3300	∥	3280–3300	⊥	~3400
C═O Stretching (Amide I)	←C—O→ ↑	1650–1660	∥	1630	⊥	1680–1700
C—N—H Bending (Amide II)	←C—N—H→ ↓	1540–1550	⊥	1520–1525	∥	<1520

Note: The amide I band at 1650 cm^{-1} involves C=O stretching (80%), C—N stretching (10%) and in-plane N—H bending (10%) contributions. The amide II band at 1550 cm^{-1} involves a mixture of C—N stretching (40%) and in-plane N—H bending (60%) contributions. Dichroism refers to polarization along the long axis.

TABLE 7.7 Characteristic vibrations localized in the side chains of proteins

Side chain	IR band (cm^{-1})	Assignment
Cys	2 600–2 700	SH stretching, very weak
Gln, Asn	32 000–3 430	NH_2 stretching
Glu, Asp	1 710–1 715	Carboxyl CO stretching
	1 570, 1 410	Ionized carboxyl COO^- stretching
Arg	1 625, 1 670–1 680	Guanidinium modes
Lys	1 600, 1 495	NH_3^+ deformation
Tyr	1 515	Mode of p-substituted benzene ring
Phe	1 495	Mode of monosubstitued benzene ring
Trp	3 400	Stretching of indole NH

However, the amide II bands are strongly overlapped by bands originating from amino acid side chains vibrations.

For oriented polypeptide chains, IR dichroism provides an informative technique. In α helix, N—H ··· O═C peptide hydrogen bonds are oriented parallel to the long axis of the molecule. Therefore the N—H stretch and amide I bands should preferentially absorb IR radiation when the direction of polarization is parallel to the helix axis. The amide II band should show the opposite trend. In β sheets, the long axis of the structure is along the extended peptide chains and the hydrogen bonds are perpendicular to this axis. Therefore each IR band in β sheets would have chroism opposite to that observed for the corresponding band in an α helix.

The frequency-correlation (Table 7.6) provides a useful empirical guide for the conformation analysis of proteins. Since the correlation is derived theoretically from α-helical structures of infinitive length and β-pleated structures of infinitive sheets, its application to the short and/or irregular α-helix and β-sheet sections in globular proteins cannot be assumed to follow the same frequencies and intensity distribution as long chains. Furthermore, considerable overlap between the various components may complicate the interpretation of the correlation. The analysis of the unknown protein may be successful only as long as the spectral features characteristic of this protein can be recognized in the spectra of calibration proteins.

The vibrational transitions of protein side-chain groups are highly localized (Table 7.7), therefore they can be applied directly to investigate the side chains of peptides and proteins (Singh, 2000).

TABLE 7.8 Characteristic IR bands of carbohydrate structures

Assignment	IR band (cm^{-1})
Ring vibration (asymmetric)	917 ± 13
Ring vibration (symmetric)	770 ± 14
Anomeric C—H (equatorial)	844 ± 8
Anomeric C—H (axial)	891 ± 7
C—H (equatorial)	880 ± 8
C—O—C stretching	1160
O—H stretching (primary)	3642
O—H stretching (secondary)	3629

In nucleic acids, most attention has been focused on the IR spectral region 1500–1800 cm^{-1}, which contains vibrations of the carbonyl and the double bonds of the purine and pyrimidine rings. The vibrations are highly sensitive to base pairing because the atoms involved participate directly in the formation of hydrogen bonds. Furthermore, each of the four common bases has a distinct IR spectrum in this spectral region (Thomas, 1969). Thus it is possible to examine A-T(U) and G-C pairs separately.

The IR spectra of carbohydrates present several characteristic features from which functional group assignments can be made. There is a strong band at 3500 cm^{-1} due to the hydrogen bonded O—H groups. The region of 1150–1000 cm^{-1} has several closely spaced vibration bands, probably arising from C—C and C—O vibrations. At lower frequencies, different structural isomers are shown to have different characteristic absorption bands, one of which is the anomeric bands. α Anomers (α-glycosides and glycoses) show a distinct absorption band at approximately 844 cm^{-1} and β anomers at 891 cm^{-1} (Table 7.8). The vibration band observed at around 875 cm^{-1} is probably due to C—O—C stretching of the pyranose ring. Ring vibrations of pyranoses occur at 917 and 770 cm^{-1} while furanose rings exhibit absorption at 924 and 799 cm^{-1}. 2-Keto glycoses show characteristic absorption at 810 and 874 cm^{-1} characteristic of the ketal group, regardless of whether it is a furanose or pyranose ring.

The IR spectrum offers a means for determining anomeric configuration and is useful in studies of polysaccharides with homologous anomeric linkages, since the anomeric C—H bonds are not influenced by the types (glycose units) and positions of linkages. An estimate of the strength of hydrogen bonds can be made from the separation in wave number ($\Delta\nu$) between the IR bands of free versus hydrogen bonded hydroxyl groups. In general, the smaller the length of the hydrogen bond, the stronger the bond itself and the greater the value of the $\Delta\nu$. Studies indicate that the crystalline region of cellulose is completely hydrogen bonded and whereas amorphous region is not.

IR dichroism is a useful technique for investigating conformations or oriented polysaccharide samples. For example, cellulose I possesses a unit cell containing two parallel chains of repeating cellobiose units. Dichroism studies show that the polarization of one of the bands is perpendicular, involved in inter-chain hydrogen bonds and two are parallel to the long axis forming intra-chain hydrogen bonds.

7.5 NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY

7.5.1 Basic principles

Nuclear magnetic resonance (NMR) is the spectroscopic method used to observe nuclear-spin reorientation in an applied magnetic field. Various applications of NMR in biochemistry include structural identification of biomolecules, chemistry of individual groups in macromolecules, structural and dynamic information of biomacromolecules, metabolic studies, as well as kinetic and association constants of ligand bindings to macromolecules (Evans, 1995; Roberts, 1993; Stassinopoulou, 1994). The technique monitors the absorption of energy associated with transition of nuclei between two nuclear magnetic energy levels. The NMR phenomenon is observable because certain nuclei behave like tiny spinning bar magnets. The spinning nucleus generates a magnetic field and thus has an associated magnetic moment, which interacts with the applied field. Most important among such nuclei in biomacromolecular applications are ^{1}H and ^{13}C having nuclear spin values (*I*) of 1/2 (Table 7.9). The corresponding NMR, namely proton magnetic resonance (PMR) and ^{13}C-magnetic resonance (CMR) will be considered in this section.

The sensitivity of the ^{13}C nucleus is 1.6% that of ^{1}H for equal numbers of nuclei in the same magnetic field. By taking into account of its natural abundance of only 1.1%, this reduces the relative sensitivity of ^{13}C as being roughly 1.8×10^{-4} that of ^{1}H. However, CMR has several advantages. First, in most cases, the chemical shifts of ^{13}C occur over a much broader range than chemical shifts of ^{1}H. The large range of ^{13}C chemical shifts permits separate visualization of many individual nuclei in various monomers of biomacromolecules. Second, CMR spectra are expected to be first-order, i.e. no ^{13}C-^{13}C spin–spin couplings would normally be detected because ^{13}C is only 1.1% abundant so that most neighboring nuclei are nonmagnetic ^{12}C. Third, ^{13}C can be inserted into a molecule at a specific locus to replace ^{12}C. In this way, a site-specific magnetic probe is obtained that does not perturb the structure. However, ^{13}C shifts are much less sensitive to the environment than are ^{1}H shifts. Therefore it is more difficult to explore conformational changes and effects of ligand interactions.

Nuclei with a net magnetic dipole such as ^{1}H and ^{13}C will orient the dipole axis in an external magnetic field in certain quantized orientations. The number of possible orientations is given by $2I + 1$. If a nucleus with $I = 1/2$ is placed in a uniform magnetic field, it may take up one of two orientations with respect to the field (the external magnetic field H defines the z-axis). Those may be considered as a low-energy orientation in which the nuclear magnet is aligned with the field (having quantum numbers $m_s = +1/2$), and those referred to as a high-energy orientation in which the magnet is aligned against the field (having quantum numbers $m_s = -1/2$). The transition between these two energy states can be brought about by the absorption of suitable electromagnetic radiation of energy.

The nuclear magnetization has direction (i.e. vector). The component, M_z, which is defined to be along the applied magnetic field (H_o) direction, and the components M_x and M_y, at right angles to H_o. The M_{xy} components arise because the spins do not align perfectly along H_o. At equilibrium, the spins are randomly distributed and the net $M_{xy} = 0$. On application of a rotating radiofrequency field with frequency at or near precession frequency, the

TABLE 7.9 Nuclei ($I = ½$) of importance in biochemical NMR

Nucleus	Relative NMR frequency	Gyromagnetic ratio (10^7 rad/s·T)	Relative sensitivity	Natural abundance (%)
^{1}H	100.000	26.752	1.00	99.98
^{13}C	25.144	6.728	1.59×10^{-2}	1.11
^{15}N	10.133	−2.713	1.04×10^{-3}	0.37
^{19}F	94.077	25.182	8.30×10^{-1}	100
^{31}P	40.481	10.839	6.63×10^{-2}	100

spins resonate and the random distribution changes into a coherent net M_{xy} component. Spin systems giving rise to net M_{xy} components are known to be phase coherent.

Nuclear spins with different electronic environments may be brought into resonance by either one of two techniques. In the frequency-sweep method, the spectrum is recorded by sweeping the applied radiation frequency. In the transient-response method, the transient signal for its component frequencies is sorted/transformed after an induction (by pulse(s) of an applied field in terms of angles, e.g. 90° or 180°) of transient response in the system. The transient signal is changed into a normal spectrum by Fourier transformation in the transient-response method, which is used by most modern NMR spectrometers. There are four parameters that define the NMR spectrum:

7.5.1.1 Intensity or area under resonance peak. The amplitudes of the resonance/transient signals are directly proportional to the number (concentration) of nuclei in an equilibrium system of nuclear spins. Relative concentrations are usually measured from the resonance intensities (peak heights). This provides an indication for the relative quantities of the resonance nuclei with different chemical environments (chemical structures) in the molecule.

7.5.1.2 Chemical shift. The frequency (ν) at which any nucleus will resonate in the NMR spectrum is give by

$$\nu = \gamma H/2\pi$$

where γ is a constant known as the magnetogyric ratio and H is the local field experienced by the nucleus, which corresponds to the applied magnetic field (H_o) by $H = H_o(1 - \sigma)$ since the nucleus will usually be shielded by the surrounding electrons. The extent of this shielding is represented by a shielding parameter (σ). It becomes

$$\nu = \{\gamma H_o(1 - \sigma)\}/2\pi$$

indicating that nuclei with different shielding parameters, i.e. different electronic environments may be brought into resonance either by a frequency sweep or by a field sweep of the spectrum. In principle, the NMR spectrum can be recorded either in cycles per second (hertz, frequency units) or milligauss (field units), since the frequency of any given resonance will be proportional to the applied magnetic field. Available spectrometers may employ different field strengths, which have been increased with improved spectrometers. A scale of field-independent, frequency-referenced units, called the chemical shift scale is chosen. The chemical shift (δ) is defined as

$$\delta = (\nu_{ref} - \nu_s)\ (\mathrm{Hz})/\nu_o\ (\mathrm{MHz}) = 10^6[(\delta_{ref} - \delta_{obs})/\delta_{ref}]$$

where ν_{ref}, ν_s and ν_o are resonance frequencies of a reference compound, sample in hertz (Hz) and operating frequency in MHz respectively. δ_{ref} and δ_{obs} are the positions (in Hz) for a reference compound and the signal of interest. The numerator is expressed in Hz as opposed to the denominator, which is expressed in MHz and therefore the unit of chemical shift parameter, δ, is expressed in parts per million (ppm). This permits us to compare spectra obtained on instruments operating at different field strengths since the frequency of a particular resonance increases in direct proportion to the increase in the field strength. Usually a compound with shield nuclei that resonate at particularly high frequencies is used as the reference. For example, tetramethylsilane (TMS) or sodium 2,2-methyl-2-silapentane-5-sulfonate (DSS) are commonly used to calibrate the instrument in proton magnetic resonance (PMR). In some literatures (especially organic chemistry), τ-scale of chemicals shifts where $\tau = 10 - \delta$ is used.

Two classes of shift are:

1. primary or intrinsic shift which is characteristic of a particular chemical group; and
2. secondary or induced shift arising from the influence, through space, of neighboring magnetic centers.

The primary shifts vary widely with different nuclei of different groups and are sensitive to the ionization state of the molecule.

The primary shifts depend on the electron cloud around the proton that is undergoing resonance. Electronegative substituents withdraw electron density from the proton, giving less shielding and therefore larger δ values. The secondary (induced) shifts are important in macromolecular spectra. NMR spectra of native (folded) biopolymers are generally different from those of the denatured (unfolded) biopolymers. An analysis of the difference spectra provides useful information about the conformation of the biopolymers. Interatomic shielding forming neighboring atoms can augment or oppose the applied field, and therefore the shift of the resonating nuclei. A special kind of interatomic shielding is ring current (delocalized π electrons) shifts of aromatic rings that are particularly important in biomacromolecules. The induced electron currents create large magnetic field affecting (positively or negatively) the resonating nucleus, depending on its position relative to the aromatic ring. The resonating nucleus on the side of the ring experiences an augmentation of the applied field while that over the center of the ring sees an opposing effect. Hydrogen bonding affects the bonded proton by causing a downfield shift relative to the unbonded state.

7.5.1.3 Multiplicity and spin–spin coupling. Spin–spin interactions (couplings) between magnetic nuclei are responsible for multiplet structure of NMR spectra. These interactions are communicated between the nuclei by electrons in a chemical bond. Thus spin–spin interactions are often referred to as through-bond interactions. The splitting (of multiplet) is independent of the field, and depends only on the nature of the bonding between the two groups, thus it is informative of the neighboring groups. The size of the interaction is defined by the spin–spin coupling constant (J), which measures the splitting between two different nuclei in a given molecule in hertz (Hz).

For the spin–spin coupling to systems containing n bond-sharing protons, the following rules can be formulated:

1. If a proton has a neighbors, sets n_a, n_b, n_c . . . of chemically equivalent protons, the multiplicity of its resonance will be $(n_a + 1)(n_b + 1)(n_c + 1)$. . . .
2. For one neighboring group of n equivalent protons, the relative intensities of the n + 1 multiplet components are given by the coefficients of the terms in the expansion of $(x + 1)^n$.

These two generalizations form the basis for interpreting spin–spin coupling patterns in NMR spectra by the first-order approximation, if J is smaller than the chemical shift frequency $\Delta\nu$ between adjacent groups (typically $J \leq \Delta\nu$). If $J > \Delta\nu$, the spectrum can be more complex than the first-order splitting pattern. An increase in field strength raises $\Delta\nu$ (chemical shifts are independent of field strength) to the point where simple first-order splitting can be seen ($J \leq \Delta\nu$).

The appearance of a multiplet also depends on the relative magnitudes of δ and J for the coupled nuclei. If $\Delta\delta >> J$, then a well-splitted multiplet is observed. However, the nuclei are said to be equivalent and no multiplet structure is observed if $\Delta\delta \approx J$. The spin–spin coupling between protons three bonds apart depends on the relative orientation of the nuclei involved in the coupling according to the Karplus equation (Karplus, 1959):

$$^{x}J_{yz} = A\cos^2\theta + B\cos\theta + C$$

For protons, θ is the angle between the protons for germinal (two-bond) coupling (x = 2) or the dihedral angle between the protons for vicinal (three-bond) coupling. Subscripts, y and z define the nuclei that are interacting. A, B and C are empirical constants, e.g. for vicinal couplings (*$\theta = \phi - 60°$ for proteins):

y	z	A	B	C
HC	CH	17	0	1.1
HC	NH	12	0	0.2
HC	OH	10	0	−1.0
HN	$C_\alpha H^*$	6.4	−1.4	1.9

The splitting from spin–spin coupling can be removed if the relative orientation of coupled nuclei changes relatively rapidly compared with $1/J$ or by irradiation of one of the nuclei or groups of nuclei with a selective radiofrequency field known as the double resonance technique. The resulting collapse of the spin multiplet is called decoupling. For a complicated molecule, the double resonance technique is used to determine which peaks belong to nuclei that are close enough to interact with another nucleus through a bond.

7.5.1.4 Relaxation times. NMR instrumentation uses an oscillating magnetic field to excite nuclei and make them precess in phase. However, without the oscillating magnetic field they will eventually lose phase coherence and also lose excitation energy. Two relaxation processes are involved in these losses.

In spin-lattice relaxation, a nucleus in the excited state interacts with fluctuating magnetic fields generated by any other atoms in the molecule (lattice). This interaction is an enthalpic relaxation because it generates heat by bringing the nucleus back to the ground state. The relaxation time (T_1) for the fraction $1 - 1/e$ of the nuclei to relax via this process can be measured by saturating the sample with a strong magnetic field over a period of time, so that the number of nuclei in the ground state equals the number of nuclei in the excited state (no phase coherence and absorption). The recovery to equilibrium can then be monitored. In practice, the measurement of T_1 is carried out by a $180° - \tau - 90°$ pulse sequence (τ is a time interval between the two pulses allowing for mixing of spins). The 180° pulse inverts the populations so that the magnetization component along the *z*-axis, M_0 becomes $-M_0$ following the recovery along the *z*-axis, and M_z with a first-order rate constant $1/T_1$ according to

$$M_z(t) = M_0(1 - 2e^{-t/T1})$$

However, the recovery taking place along the *z*-axis cannot be observed directly because there are no M_{xy} components (no phase coherence). Therefore the recovery is monitored by applying a 90° pulse at intervals, thus tipping the resultant *z*-magnetization into the *xy*-plane.

The spin–spin relaxation process involves dipole–dipole interaction between an excited state spin and a ground state spin for the same type of nucleus. The interaction leads to an exchange of energy making the system more random (entropic relaxation), and eroding the phase coherence. The spin–spin relaxation does not change the net population of the excited state. The relaxation time (T_2), which is defined as the time constant of the decay of the M_{xy} components, is the time for 1/e of the nuclei to relax via this process. T_2 is related to the line-width (half-width at half-height) by the following relationship:

$$\Delta\nu_{1/2} \approx 1/(\pi T_2).$$

Since T_2 is related to line-width for freely tumbling molecules in solution, the dynamic state of the resonance nucleus can be inferred from its T_2.

An important effect that is observed in NMR spectroscopy is the nuclear Overhauser effect (NOE). This is a through space interaction where a change in signal intensity for one type of nucleus is caused by irradiation of another type of nucleus that is nearby. This change in intensity results from the T_1 relaxation of one of the pair of spins caused by the spin being irradiated. This causes changes in the ratio of the ground state population to the excited state population thus changing the intensity. The NOE is a dipole–dipole interaction that depends strongly on the distance between the two types of nuclei by r^{-6}, and can be used to measure interatomic distances up to 0.5 nm.

The basic strategy for biomolecular structure determinations by NMR follows three stages:

1. sequence-specific assignment of backbone and side-chain resonances using experiments that demonstrate through-bond (scalar) and through-space (<5 Å) connectivities;
2. identification of as many through-space NOE connectivities as possible, which yield a large set of approximate interproton distance restraints as well as determination of torsion angle restraints based on coupling constant data; and
3. calculation of 3D structures on the basis of these distance and angular restraints.

For these objectives, the limitations of one-dimensional (1D) NMR can be overcome by extending the measurements into a second dimension.

7.5.2 Two-dimensional Fourier transform NMR

Common NMR spectrometers are Fourier transform (FT) instruments that send a pulse of energy to excite all the transitions at once. The pulse has a nominal frequency, ν_0 and a short finite length, τ in which the short pulse still contains hundreds of wavelengths of each frequency, $\Delta\nu = \nu_0 - \nu$. The excess spins tend to precess in phase, although at different $\Delta\nu$, depending upon the frequency required for excitation of that particular type of nucleus. Each type of precessing nucleus induces an oscillating current, which can be used for detection after the pulse. However, the magnitude of the induced current will decrease with time because of the relaxation processes. The exponential decrease in the induced current with time is normally dominated by the faster spin–spin relaxation (T_2). The oscillating signal with its exponential decay is called free induction decay (FID). The frequency of the FID is the frequency of the radiation absorbed by the particular type of equivalent nuclei. The magnitude of the FID at $t = 0$ is a measure of the intensity of the absorption. The exponential decay of the FID gives the relaxation time (dominated by T_2), whose reciprocal $\Delta\nu$ is the half-width of the resonance peak at half height. Thus the FID can be transformed through a Fourier transformation (from the time domain to the frequency domain) to a NMR peak.

A pulse of just the right energy to excite half the excess nuclei in the ground state puts the net magnetization in the *xy*-plane. The coherent oscillation of the magnetization produces a maximum signal, and is called a 90° pulse (tipping the magnetization 90°). Twice the energy of a 90° pulse puts the net magnetization in the $-z$ direction and is called a 180° pulse. A sequence of pulses is used to generate 2D NMR.

2D NMR spreads the severely overlapping 1D NMR spectrum of a biomacromolecule into two orthogonal frequency dimensions. The resulting improvement in resolution leads to the popularity of 2D NMR in structural studies of biomolecules (Bax, 1989; Varani and Tinoco, 1991). With its short pulse of energy that contains all frequencies of interest, FT NMR is rapid and convenient for compiling repeated scans that can be averaged to reduce noise. The use of such pulses simplifies spectra by looking only at the interaction between nuclei and spreading these interactions over two or more dimensions. The two important interactions are spin–spin interactions (J splittings) through the bonds between nearby nuclei (J-coupled nuclei two or three chemical bonds away) and the NOE, which changes the signal intensity of one nucleus when another is irradiated nearby. Spin–spin interactions provide information about the bonding structure and NOE is used to assign resonances to their order in the sequence of a biomacromolecule, and to measure the distance between nuclei. Since NOE is a dipole–dipole interaction through space, it is sensitive to nuclei that are close together by bonding structures as well as to nuclei that are close together (~0.5 nm) due to the conformational structure of the molecule. The most convenient 2D method used to measure J-coupling is J correlated spectroscopy (COSY) and that for measuring NOE is called the nuclear Overhauser effect and exchange spectroscopy (NOESY). The pulse sequences consist of a 90° pulse, an evolution period for the 2D, a 90° pulse for COSY, a 90° pulse (initiating evolution period), a second 90° pulse (starting mixing period) and a third 90° pulse (preceding detection of FID) for NOESY.

Usually the 2D NMR spectrum is either for the spin–spin interaction (COSY) or for the NOE (NOESY). Spectra of 2D NMR are plotted on two frequency axes, which show a crosspeak at the frequencies corresponding to each pair of nuclei that interact. The magnitude of the interaction effect is represented by the intensity of the crosspeak. It is shown by the number of contours (like a contour map) at the point of interaction; the more contours, the higher the peak corresponding to the interaction. The diagonal represents 1D spectrum of the contour line of peaks, which also plot along both axes.

Protocols for obtaining resonance assignment for 2D NMR of proteins follow two steps. The first step in resonance assignment uses J-coupling (through-bond) information provided by the COSY experiment (and other advanced techniques) to classify resonances according to the types of corresponding monomers. For example, Gly is the only residue where two protons interact with amide proton, Ala is the only residue where the $C_\alpha H$ interacts with two C_β methylene protons that do not split, and Pro is the only residue without amide hydrogen. The second step concerns the identification of the sequence specific position of each monomer unit. NOE interactions between adjacent monomer units are sought for unique dimer segments in the biomacromolecular backbone (from the known sequence). Such unique dimers provide a starting point for a further sequential resonance assignment based on the NOESY experiment.

7.5.3 NMR of proteins (Wüthrich, 1986)

Amino acids have characteristic resonances resulting from the C_α-proton and the protons on the side chains.

Proteins contain many different types of protons (Figure 7.2), therefore their 1D PMR spectra are complex. The peptide NH resonates around 8–9 ppm.

The PMR spectra of side chains can be divided approximately into four regions:

1. The lowest field peaks (7–10 ppm) are due to aromatic protons of Phe, Tyr and Trp and ring NH of His and Trp.

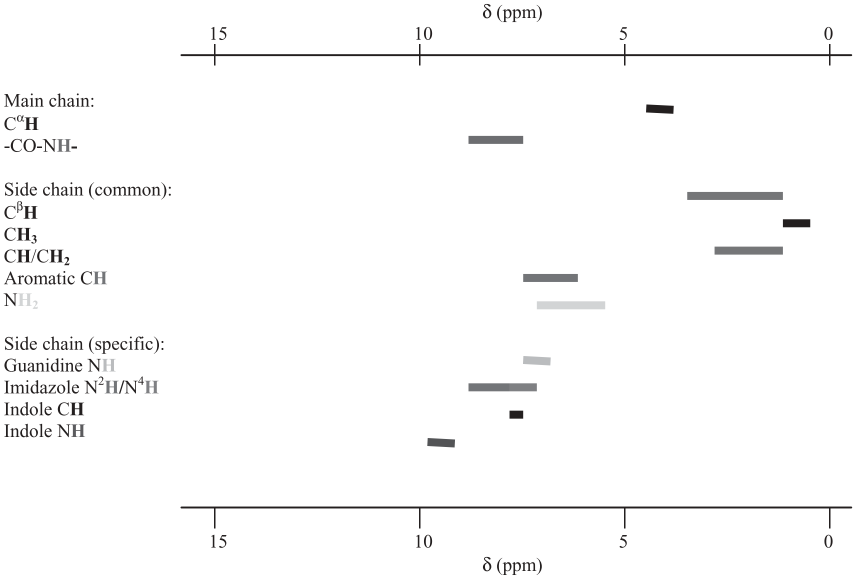


Figure 7.2 Ranges for ^1H (bold) chemical shifts of amino acid residues in coil conformation at pH7.0
The characteristic chemical shifts for proton (shown in bold) resonances with respect to a reference compound, R—Si(CH_3)$_3$ ($\delta = 0$ ppm) are illustrated.

2. Proceeding upfield (~4 ppm) to the second peaks, which are assigned to α-CH protons of all amino acids, the β-CH protons of Thr and Ser, and CH_2 protons of Arg.
3. Proceeding upfield to the third peaks, which comprise the CH_2 groups of Cys, Met, Asp, Lys, Pro, His, Phe and Tyr.
4. The highest field peaks are the overlapping of two broad peaks. One of the peaks comprises the aliphatic CH group of Leu, Ile and Val, the CH_2 group of Met. The second peak (higher field) arises from the CH_3 groups of Ala, Val, Leu, Ile and Thr.

The spectral changes accompanying helix formation in proteins can be generalized as:

- The α-CH peak shows an upfield shift on helix formation, accounting to 0.3–0.6 ppm.
- The NH resonance shifts downfield by about 0.2 ppm on helix formation.
- The NH and α-CH resonances are markedly broadened on helix formation.
- The resonances of side-chain protons show no change in chemical shift on helix formation. Any such change in proteins must therefore be ascribed to the tertiary structure.

The chemical shifts of discriminatory nuclei for amino acids in random coil are summarized in Table 7.10.

The NOESY experiments that yield 2D spectra show correlation for pairs of protons that are in close proximity of each other, mostly for protons that are less than five amino acid apart (short range) in the peptide sequence. Short-range distances (<2.5 Å) invariably yield substantial NOEs, medium-range distances (2.5–3.5 Å) yield weaker NOEs, while

TABLE 7.10 Chemical shifts (ppm) for amino acids in random coil

	$^{1}H_{\alpha}$	$^{1}H_{N}$	^{15}N	$^{13}C_{\alpha}$	$^{13}C_{\beta}$	$^{13}C_{O}$
Gly	3.96	8.33	108.8	45.1		174.9
Ala	4.35	8.25	111.2	61.9	40.8	175.6
Val	4.12	8.03	119.2	62.2	32.9	176.3
Leu	4.34	8.16	121.8	55.1	42.4	177.6
Ile	4.17	8.00	119.9	61.1	38.8	176.4
Pro	4.42			63.3	32.1	177.3
Ser	4.47	8.31	115.7	58.3	63.8	174.6
Thr	4.35	8.15	113.6	61.8	69.8	174.7
Cys (red)	4.55	8.32	118.8	58.2	28.0	174.6
Cys (oxid)	4.71	8.43	118.6	55.4	41.1	174.6
Met	4.48	8.28	119.6	55.4	32.9	176.3
Asp	4.64	8.34	120.4	54.2	41.1	176.3
Glu	4.35	8.42	120.2	56.6	29.9	176.6
Asn	4.74	8.40	118.7	53.1	38.9	175.2
Gln	4.34	8.32	119.8	55.7	29.4	176.0
Lys	4.32	8.29	120.4	56.2	33.1	176.6
Arg	4.34	8.23	120.5	56.0	30.9	176.3
His	4.73	8.42	118.2	55.0	29.0	174.1
Phe	4.62	8.30	120.3	57.7	39.6	175.8
Tyr	4.62	8.30	120.3	57.7	39.6	175.8
Trp	4.66	8.25	121.3	57.5	29.6	176.1

Notes: 1. Chemical shifts (in ppm) at 25°C are taken in part from Wishart and Nip (1998).
2. The values for Phe are taken from Tyr.

NOEs of protons with distance greater than 3.5 Å are often too weak to be observed although their presence, if detectable, may provide useful information about the chain folding. Distances between backbone ($C_{\alpha}H$, NH) protons and between a backbone and $C_{\beta}H$ protons on adjacent amino acids are termed sequential distances. which are strongly dependant on the protein structure.

In the allowed region of the conformation map, it is noted that at least one sequential distance is always less than 3 Å, capable of yielding NOE correlation. The two major secondary structures of proteins show recognizable fingerprints in the NOESY spectrum. α-Helices are characterized by a sequence of short sequential NH-NH connectivities, corresponding to the 2.8 Å interproton distance (d_{NN}). Interresidue NH-$C_{\alpha}H$ distances are also short (2.4 Å) and give rise to intense correlation. Weak interresidue correlation are often observed between NH and the $C_{\alpha}H$ of the preceding residue ($d_{\alpha N}$ connectivity) and to the $C_{\alpha}H$ of the residue three positions ahead in the sequence ($d_{\alpha N(i,i+3)}$). Correlations between NH resonances and the $C_{\beta}H$ of the preceding residue ($d_{\beta N}$), as well as weaker $d_{\alpha N(i,i+2)}$ and $d_{\alpha N(i,i+4)}$ connectivities, are also commonly observed. Antiparallel β-sheets are characterized by intense $C_{\alpha}H$-$C_{\alpha}H$ cross peaks for residues on opposite strands, in addition to an intense sequential $d_{\alpha N}$ connectivity. Parallel β-sheets show the strong sequential $d_{\alpha N}$ connectivity, but with weak $C_{\alpha}H$-$C_{\alpha}H$ cross peaks.

NMR spectroscopy has been applied as an alternative approach to X-ray crystallography, to elucidate structures of proteins (Tugrinov *et al.*, 2004). The database for NMR spectra of biomacromolecules, mainly proteins, can be accessed at BioMagResBank (BMRB), http://www.bmrb.wisc.edu/page/homeinfo.html.

7.5.4 NMR of nucleic acids (Wüthrich, 1986)

The nucleotide monomers that make up a nucleic acid strand consist of a phosphate with no protons, a pentose with C′-protons, primary and secondary hydroxyl protons, and protons of purine and pyrimidine rings. The PMR spectrum of a nucleic acid approximates the superposition of the PMR for the constituent nucleotides (Figure 7.3):

- The methyl resonance of T occurs at 1.4 ± 0.2 ppm.
- The resonance peaks of pentose C′-H appear at 2.2 ± 0.4 ppm (C_2'-H), 3.8 ± 0.4 ppm (C_3'-H and C_5'-H), and 5.4 ± 0.8 ppm (anomeric H).
- Protons of pyrimidine rings resonate at 5.6 ± 0.6 ppm (C_5-H) and 7.2 ± 0.4 ppm (C_6-H).
- Protons of purine rings resonate at 7.8 ± 0.6 ppm (C_2-H and C_8-H).
- The amino resonance peaks of A, G and C (exocyclic N-H_2) occur at 7.7 ± 1.0 ppm.
- The cyclic imino protons (ring N-H) of pyrimidine and purine rings resonate at 12.5 ± 2.5 ppm.

High resolution PMR in H_2O provides a powerful tool for investigating hydrogen bonded base pairing structures in nucleic acids. The resonances observed in the low field region extending from 11–15 ppm, are ascribed to the hydrogen bonded ring NH protons. Since each base pair contains just one ring NH proton (imino proton), i.e. U(T)-N_3**H** or G-N_1**H**, each proton detected corresponds to one base pair (bp) in the molecule. For example, the hydrogen bonded (base paired) ring NH protons are shifted upfield from their standard position (i.e. $(AU)^0 = 14.7 \pm 0.1$ ppm and $(GC)^0 = 13.6 \pm 0.1$ ppm) by ring current fields from their nearest neighbors. Imino proton resonances in RNA helices, which are in A-form, are easy to assign because the imono protons in successive bps are close enough to give through-space, i.e. NOE correlations. U-N_3**H** in AU pairs (13–15 ppm) gives intense NOEs to the AH_2 proton resonances of their hydrogen-bonding partners. G-N_1**H** protons of GC pairs (12–14 ppm) give strong NOEs to the CN_4 protons of their hydrogen-bonding partners and weaker NOEs to CH_5's. Furthermore, the base type of an imino proton resonance can be determined by experiments that correlate imino proton resonance with the ^{15}N resonance of the nitrogen's to which they are bonded (Moore, 1995).

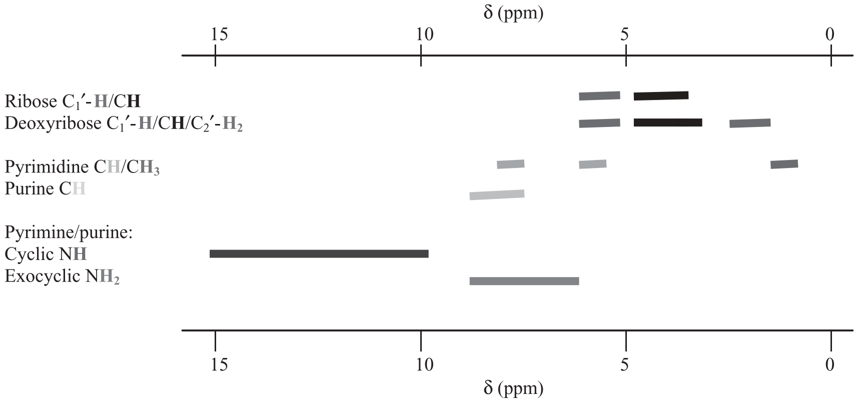


Figure 7.3 Ranges for ^1H (bold) chemical shifts of nucleotides

7.5.5 NMR of glycans (Jimenez-Barbero and Peters, 2002)

The application of NMR in structural analyses of oligo- and polysaccharides includes:

1. Identification of the component sugar residues:
   a) establishing the number of constituent sugar residues;
   b) establishing the types of constituent sugar residues and their anomeric configuration.
2. Determination of saccharide sequences and interresidue linkage positions;
3. Extraction of conformational information.

The PMR spectrum of an oligo-/polysaccharide approximates the superposition of those of its constituent monosaccharides (Figure 7.4).

The interpretation of PMR spectra for carbohydrates has led to the formulation of several empirical rules:

- Axial protons tend to resonate at higher field strength than do equatorial protons.
- The coupling constants between axial protons on adjacent carbons (5–8 cycles per s) are two to three times larger than those observed between axial-equatorial or diequatorial coupling constants.
- In general, axial and equatorial protons on the same carbon will be strongly coupled with $J \approx 12$ cps.
- The anomeric proton absorbs at a unique field strength (at lower field strength than do any of the other ring protons) and its coupling with the C-2 proton, e.g. $\delta = 5.26$ ppm ($J_{1,2} = 3.0$) for α-D-glycoses and $\delta = 4.66$ ppm ($J_{1,2} = 7.5$) for β-D-glycoses.
- Hydrogen bonding causes a downfield shift relative to the free (unbound) proton.
- Protons on substituents such as acetoxy and methoxy groups are not coupled to other protons and thus give rise to fairly sharp spectral resonances.
- The prevailing ring form in solution can be readily be determined. For example, the resonance signal of the hydroxymethyl group attached to the ring does not occur for pentose in the pyranose ring.

The CMR data for monosaccharides (Bock and Pederson, 1983), oligosaccharides (Bock *et al.*, 1984) and polysaccharide (Gorin, 1980) have been compiled. The NMR

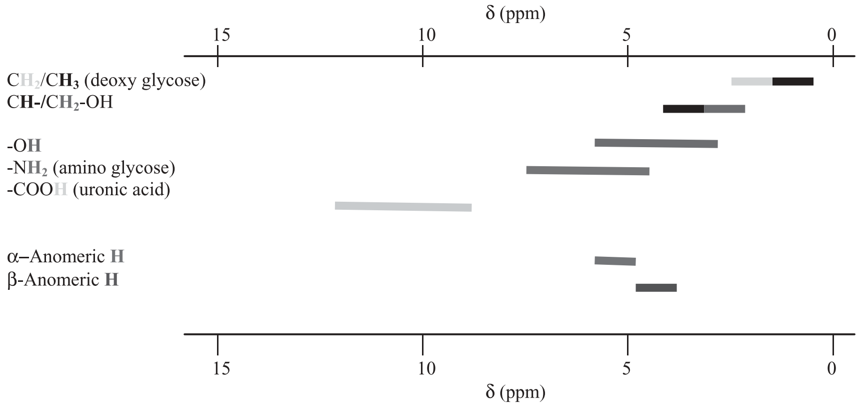


Figure 7.4 Ranges for ^1H (bold) chemical shifts of glycans

spectra of saccharides can be accessed from NMR database of Glycosciences at http://glycosciences.de/sweetbase/nmr/.

7.6 OPTICAL ROTATORY DISPERSION AND CIRCULAR DICHROISM SPECTROSCOPY

7.6.1 Basic principles

The optical activity arises from the chiral (asymmetric) centers of molecular structures. An inherently asymmetric chromophore, e.g. peptide bond, is optically active. An optical activity can also be induced in inherently symmetric chromophores from interactions with asymmetrically placed neighboring group. Optical activity is observed when a transition occurs in the asymmetric environment, e.g. nucleotides in nucleic acids. An optically active molecule interacts differentially with left- and right-circularly polarized light. This interaction can be detected either by optical rotatory dispersion (ORD) or circular dichroism (CD) (Fasman, 1996). Both ORD and CD are closely related to the absorption spectrum of the molecule under study, because it is the same electronic transitions that are involved in all three phenomena. Some of the electronic transitions may interact with their local environment and become optically active, each of these having a component CRD and CD spectrum. While electronic transitions always give rise to positive absorption bands, the ORD and CD bands can be positive or negative in sign and need not bear any relation to the size of the absorption band that give rise to them.

An interaction of a molecule with plane-polarized light, which consists of two circularly polarized beams rotating in opposite directions (left, L and right, R), affects:

1. the velocity of the beam through the sample, which is characterized by the refractive index (n_L and n_R); and
2. the absorption by the sample, which is characterized by the extinction coefficient (ε_L and ε_R).

Optical activity can be detected either as the differential refractive index ($n_L \neq n_R$) or as the differential absorption (dichroism) ($\varepsilon_L \neq \varepsilon_R$). If $n_L \neq n_R$, the rotation velocity of L and R in the sample will be different, and this phenomenon, with respect to variation in wavelengths (dispersion), gives rise to ORD. Thus the ORD spectrum records the magnitude of the optical rotation resulting from a differential change in velocity of the two beams of the polarized light characterized by the differential refractive index with wavelength. If $\varepsilon_L \neq \varepsilon_R$, the sum of beams, L and R passing through the sample, is no longer a plane-polarized beam but is elliptically polarized. The phenomenon of CD is observed. Thus CD spectrum records a differential absorption of each beam of the polarized light with wavelength. This results in a distortion of the plane-polarized beam so that the resultant electric vector of the light wave oscillates in an ellipse and CD measures the ellipticity of the light wave. ORD is characterized by $[\alpha]_\lambda$, which is the specific rotation at a given wavelength or the molar rotation $[\phi]_\lambda$. CD is characterized by ΔA (the differential absorption of the two beams) or the molar ellipticity $[\theta]_\lambda$.

In ORD, the rotation is measured as

$$[\alpha]_\lambda = \alpha_{obs}/cl$$

where $[\alpha]_\lambda$ and α_{obs} are the specific rotation at wavelength λ and the observed rotation (in degree) respectively. C is the sample concentration (g/mL) and l is the path length (dm, i.e. 1 dm = 10 cm). The molar rotation, $[\phi]_\lambda$ (degree · cm^2 · dmol^{-1}) for a sample with the molecular weight, M_w is defined according to

$$[\phi]_\lambda = ([\alpha]_\lambda M_w)/100 = (\alpha_{obs} M_w)/100cl$$

and the mean residue rotation, $[m]_\lambda$ (degree · cm^2 · dmol^{-1}) is defined as

$$[m]_\lambda = ([\alpha]_\lambda \cdot MRW)/100$$

where MRW is the mean residue weight of monomers that constitute the biomacromolecule (e.g. the molecular weight of biomacromolecule divided by the number of monomers).

Both left- and right-handed circularly polarized light obey Beer's law and CD is defined as the difference in extinction coefficient as

$$\Delta A = \Delta\varepsilon cl = A_L - A_R$$

where $\Delta\varepsilon = \varepsilon_L - \varepsilon_R$, i.e the absorbencies of the L and R beams (A_L and A_R). In CD, the ellipticity, θ used in relation to ORD, is related to $\Delta\varepsilon$ by

$$\theta = 3300\Delta\varepsilon \,\text{degree} \cdot \text{cm}^2 \cdot \text{dmol}^{-1}$$

The molar ellipticity $[\theta]_\lambda$ (degree · cm^2 · dmol^{-1}) is defined as

$$[\theta]_\lambda = (\theta_{obs} \cdot M_w)/100cl$$

and the mean residue ellipticity, $[\theta']_\lambda$ (degree · cm^2 · dmol^{-1}) is defined similarly as

$$[\theta_m]_\lambda = (\theta_{obs} \cdot MRW)/100$$

The optical rotation and ellipticity are measured over a range of wavelengths. Thus ORD and CD spectra, like absorption spectra, are constituted by the sum of spectral bands that attend the individual transitions of the molecule, i.e.:

$$\varepsilon(\lambda) = \Sigma[\varepsilon(\lambda)_{oa}]$$

$$[\phi(\lambda)] = \Sigma[\phi(\lambda)_{oa}] \quad \text{and}$$

$$[\theta(\lambda)] = \Sigma[\theta(\lambda)_{oa}]$$

These functions characterize the bands in the respective spectra that arise from a transition, i.e. from ground state to an excited state of the molecule. The ORD curve, with its minimum, maximum and inflection points is typical of dispersion phenomena. It is noted that the inflection point of the ORD, the maximum point (or minimum) of CD, and the maximum absorption of the same optically active transition all ideally fall at the same wavelength, λ_o. The term Cotton effect (image relationship between positive and negative Cotton effects) is applied to the characteristic ORD and CD curves in the region of the responsible absorption band.

Main applications of ORD or CD spectroscopy are the determination of the secondary structure of biomacromolecules and detection of their conformational changes. The on-line analysis of CD spectroscopic data can be accessed at DICHROWEB (http://public-1.cryst.bbk.ac.uk/cdweb/).

7.6.2 ORD/CD Spectra and protein secondary structures

Our understanding of ORD and CD spectroscopy as applied to the study of biomacromolecular conformation, particularly secondary structures of proteins has been largely derived from investigations of homopolypeptides with defined secondary structures (Greenfiel and Fasman, 1969). We assume that:

- the protein structure can be divided into homogeneous units of ordered secondary structure, each with component sets of a_i and λ_i for their peptide bond rotations;
- the contribution from each type of secondary structures is additive; and
- the peptide bond rotations are sole contributors to rotatory dispersion of the protein.

Then the multiterm Drude equation for the optical activity of proteins with varying conformation can be written as

$$[\mathrm{m}]_\lambda = \sum_{\mathrm{ij}}\sum x_j a_{ij}\lambda_{ij}^2/(\lambda^2 - \lambda_{ij}^2)$$

where x_j is the fractional content of the *j*th conformation, and a_{ij} and λ_{ij} correspond to the rotatory strength and wavelength of the *i*th absorption band of the *j*th conformation. It is, in principle, possible to estimate a_i and λ_i for each conformation and comparing these values with those for model systems of known structures to evaluate the content (x_j) of each conformation in the protein. Various simplified forms that eliminate the summation over all bands include one-term Drude equation:

$$[\mathrm{m}]_\lambda = \mathrm{k}\lambda_c^2/(\lambda^2 - \lambda_c{}^2)$$

and two-term Moffitt equation:

$$[\mathrm{m}]_\lambda = \mathrm{a_o}\lambda_o^2/(\lambda^2 - \lambda_0^2) + \mathrm{b_o}\lambda_o^4/(\lambda^2 - \lambda_o^2)^2$$

where k, λ_c, a_o, b_o, and λ_o are adjustable parameters. These are empirical equations representing special cases of the general multiterm Drude equation. A simple ORD curve, which is characteristic of a random coil conformation, can be approximated by the one-term Drude equation, whereas those of α-helices and β-sheets are better fitted with the two-term Moffitt equation. Thus the ORD curve for a mixture of helix, sheet and coil may be expressed as:

$$[\mathrm{m}]_\lambda = x^R a_o^R \lambda_o^2/(\lambda^2 - \lambda_0^2) + [x^\alpha b_o^\alpha + x^\beta b_o^\beta]\lambda_o^2/(\lambda^2 - \lambda_o^2) + [x^\alpha b_o^\alpha + x^\beta b_o^\beta]\lambda_o^4/(\lambda^2 - \lambda_o^2)^2$$

Using $\lambda_o = 212\,\mathrm{nm}$, some of the a_o and b_o values are:

Conformation	a_o	b_o
α-Helix	330 ± 190	−580 ± 20
β-Sheet	−810 ± 400	60 ± 30
Coil	−420 ± 180	−10 ± 20

Cotton effects (anomalous dispersions) are observed in the near-UV absorption bands and yield information on the content and conformation of the chromophoric residues. For a perfect α-helix, the observed three Cotton effects are the π→π* transition associated with a negative Cotton effect at about 207 nm, a strong positive band at about 191–198 nm, which is a composite of many bands and a negative effect derived from n→π* transition at 222–233 nm. For β-structures, the n→π* transition is associated with a small negative Cotton effect at 225 nm. The π→π* transition is split into three components. One (y) is polarized in the direction of the chain giving rise to a positive Cotton effect at 199 ± 1 nm. The second (x) is polarized in the direction of the interchain hydrogen bonds giving rise to a small positive Cotton effect at 195 ± 1 nm. The third (z) is polarized perpendicular to the plane of the sheet giving rise to large negative Cotton effect at 182–185 nm.

The optical activity of helical segments in proteins can vary considerably with helix length, conformation and nature of side chains. The Cotton effects at 198 and 207 nm are sensitive to helix geometry, while the 233 nm Cotton effect is not, therefore $[m]_{233}$ measurement gives better average helical contents than do measurements at other wavelengths. The ORD of perfect helices should be relatively independent of solvent because the tightly packed helix structure provide the uniform local environment for the $n \rightarrow \pi^*$ transition, while distorted helices will be more environment dependent.

The CD curve of perfect right-handed α-helix is characterized by a negative ellipticity at about 222 nm due to the $n \rightarrow \pi^*$ transition, and a negative and positive couplet at about 208 and 194 nm due to the parallel and perpendicular components of the $\pi \rightarrow \pi^*$ transition respectively. The calculated mean residue ellipticities at 194 nm and 208 nm are extremely sensitive whereas the ellipticity at 222 nm is less sensitive to chain length of helices. Among proteins, the CD spectra of α-helical structures show the ellipticities at negative maxima of 222–223 nm (variability of about 10% in ellipticity value) and 207–209 nm (variability of about 25% in ellipticity value). At the positive maximum near 191 nm, the variation in the ellipticity value is about 30% among proteins.

The CD curve of the antiparallel β-sheet is characterized by one band at about 198 nm with negative ellipticity at about 215 nm corresponding to absorption polarized parallel to the chains, one band at about 195 nm with positive ellipticity corresponding to absorption polarized in the direction of the hydrogen bonds between chains, and another band at about 175 nm with negative ellipticity corresponding to absorption polarized perpendicular to the plane of the sheet. The positive ellipticity band shifts to the red with increasing width of the β-sheets. The CD of the parallel β-structure is similar to the long-wavelength portion of the CD of the antiparallel β-sheet in that a negative band is adjacent to a positive band of slightly greater size. In the antiparallel β-sheet, the strong absorption band may be either red- or blue-shifted relative to the monomer depending on the width, but it is associated with a small positive CD band. Whereas in the parallel β-sheet, the major component in absorption is always blue-shifted and is associated with a large negative CD band.

By going from the α-helical to β-conformation, the wavelength maximum of $\pi \rightarrow \pi^*$ transition shifts from 195 nm to 189 nm. However, it shifts from 189 nm to 196 nm on increasing the number of strands in the antiparallel β-sheets. The $n \rightarrow \pi^*$ transition of the antiparallel β-structure has a negative effect at about 220 nm, which increases in magnitude with increasing number of residues in each strand but decreases in magnitude with increasing number of strands in the sheet. In the antiparallel sheet the strongest absorption band has a position depending on sheet width but it is always associated with a small positive CD band. In the parallel form the major absorption band is always blue-shifted relative to its large negative CD band. Furthermore, the difference in λ_{max} of the absorption and CD bands is much more sensitive to sheet width for the parallel (maximum $\Delta\lambda = 13$ nm) than the antiparallel structure (maximum $\Delta\lambda = 5$ nm).

The random-coil CD spectrum of homopolypeptides displays a small positive ellipticity at about 217 nm and a large negative ellipticity at about 198 nm. The CD spectra of random-coil or unordered homopolypeptides differ in two important regards from spectra of denatured proteins, namely:

1. there is no positive ellipticity peak at 216–218 nm; and
2. the negative ellipticity peak near 200 nm is less intense and may be red-shifted by as much as 4 nm.

The random coil conformation of a protein structure is the most difficult to evaluate. Usually the recognizable α-helix and β-sheet contributions to the optical activity are estimated and the random coil is estimated by subtraction.

Side chain optical activity has occasionally been observed as irregularities in the ORD and CD curves in the near UV (250–300 nm) where aromatic residues absorb light. They contribute the positive Cotton effect with peak and trough at 285 nm (Trp), 280 nm (Tyr) and small Cotton effects between 260–270 nm probably from Phe.

7.6.3 Empirical applications of ORD and CD

Biomacromolecules are asymmetric and therefore show optical activity that can be observed as ORD due to the difference in refractive index or as CD by the difference in absorption for the left and right circularly polarized light. The two phenomena are related, though CD is applied more commonly to investigate biopolymer structures because it monitors the effect of absorption bands one at time. In contrast, ORD measures the combined effect of all the bands that give rise to a refractive index. CD of biomacromolecules is usually measured in the region of electronic absorption and therefore known as electronic CD.

For proteins, it is the amide chromophore and the long-range order or lack of it that is responsible for the characteristic spectra of each of the secondary structural elements. A number of approaches have been employed to analyze the ORD or CD spectrum of a protein for the type and content of component secondary structures utilizing a set of basis spectra. The basis spectra are derived from the ORD/CD spectra of proteins from which the secondary structural content has been determined by X-ray crystallography. It is the solution of the linear combination of the basis spectra that gives the relative proportion for each of the secondary structural elements in the protein.

At 233 nm, ORD spectrum of α-helix has a negative Cotton effect while β-sheet and random-coil structures have an average $[m]_{233}$ value of $-2520\,\mathrm{deg\,cm^2\,dmol^{-1}}$, thus $[m]_{233}$ can be used to estimate the α-helical content by

$$[m]_{233} = -12\,700 \cdot f_\alpha - 2520\,\mathrm{deg\,cm^2\,dmol^{-1}}$$

The CD Spectra of β-sheet and random-coil conformations are isodichroic (identical ellipticity values) at 208 nm with an average ellipticity of $-4000\,\mathrm{deg\,cm^2\,dmol^{-1}}$. The observed ellipticity at 208 nm for α-helix is $-32\,600 \pm 4000\,\mathrm{deg\,cm^2\,dmol^{-1}}$. Thus the α-helical content (fraction of α-helical structure, f_α) of proteins can be estimated according to

$$[\theta]_{208} = -32\,600 \cdot f_\alpha - 4000\,\mathrm{deg\,cm^2\,dmol^{-1}}$$

The estimation is improved by comparing ORD/CD data with proteins of known structures. If the assumptions: i) only three major secondary structures, namely helical, β-pleated and random-coil structures contribute to protein ORD/CD; and ii) the contributions to the protein ORD/CD from these secondary structures are additive, then the rotation/ellipticity contributions of a protein can be represented by

$$[m]_\lambda = x_H f_H + x_\beta f_\beta + x_R f_R$$

$$[\theta]_\lambda = y_H f_H + y_\beta f_\beta + y_R f_R$$

where $f_H + f_\beta + f_R = 1$, and $f_H, f_\beta, f_R \geq 0$ are the fractions of the helical, β-pleated and random-coil structures respectively. The x_H, x_β, x_R, y_H, y_β and y_R are reference parameters that would be obtained if the protein molecules are made up of segments of pure helical, β-pleated and random-coil structures. These x and y parameters can be reduced from proteins whose f values are available from their known 3D structures. The resulting parameters, x_H, x_β, and x_R or y_H, y_β and y_R can then be used to determine f_H, f_β and f_R from the ORD or CD spectrum of a protein under study.

The main applications of ORD and CD have been for the estimation of the types and contents of secondary structures, following conformational changes of proteins under a variety of conditions, and probing of specific environment in protein molecules. However, the estimated content of secondary structures is often unreliable. The real power of ORD/CD is in the analysis of structural changes in a protein or in the comparison of the structure of an engineered protein or protein model to the parent structure. CD is unique in that the information about the structure of the target molecule can be determined quickly and efficiently. The information content is low, but it is usually consistent. X-ray crystallography and NMR are much more powerful tools for structural analysis, but do not lend themselves to the analysis of the structure of a large number of proteins that can easily/quickly be generated in mutational analysis or a protein design program at present time. CD can be used to analyze a number of protein samples from which interesting candidates can be selected for more detailed structural analysis.

Nucleic acids are studied by ORD and CD with a view to determining base stacking characteristics. For nucleic acids, unlike proteins, the spectral differences between different monomeric residues cannot be ignored. The nitrogen bases themselves are directly involved in close interactions in all common secondary structures. Not only the base composition but some actual sequence information must be taken into consideration to explain CD spectra of nucleic acids. The base-stacked conformation of a helical structure results in intense CD, which is characteristics of nucleic acids. The CD of a polynucleotide is the contributions from its constituent mononucleotides and adjacent base–base interactions. The positive and negative CD couplet at about 280 nm and 240 nm and the intense positive ellipticity at about 190 nm are characteristic of the B-DNA. The intense positive band at 270 nm coupled with the negative band at 210 nm, and the extremely intense positive CD at 185 nm are characteristic of A-DNA and double-stranded RNA (A-duplex). The Z-form of DNA has a negative band at 290 nm, but its characteristic CD is an intense negative ellipticity at about 195 nm.

Both theoretical and experimental observation indicated that the CD of the B-DNA is conservative, i.e. nearly similar magnitudes of the 275–280 nm positive band and 240–245 nm negative band with a crossover point near the absorption maximum of 260 nm. Whereas the CD of the A-DNA and RNA are nonconservative, i.e. unequal magnitude of the positive band versus the negative band. The Z-DNA spectrum shows an inversion of the peaks relative to the spectrum of B-DNA (Pohl and Jovin, 1972). Conformational changes from the B-form can be detected by the transition of CD spectra from the conservative to nonconservative. Such transitions are observed for complexation of DNA to histones and nucleic acids in nonaqueous solutions. CD is also convenient for monitoring the conformational changes of a nucleic acid as a function of binding, temperature and solvent.

Carbohydrates do not contain chromophores that absorb at a wavelength accessible to commercial instrumentation (the hydroxyl functional group has an absorption band at about 165 nm, which is too far down for practical studies). Although glycoses give plain dispersion curves, one immediate result is that D- and L-sugars give curves that are similar in shape but opposite in sign. Another useful parameter is the tremendously enhanced magnitude of the optical rotation at the lower wavelengths. The only aldohexoses examined, which do not show plain dispersion curves, are D-galactose and D-talose. Carbohydrates that have a significant free amount of free aldehyde form in solution might exhibit Cotton effects above 200 nm where $n \rightarrow \pi^*$ transition would be expected to occur. It is also noted that α and β glycosides of a given pyranose exhibit different shape of their dispersion curves.

7.7 X-RAY DIFFRACTION SPECTROSCOPY

7.7.1 Basic principles

The method that currently yields reliable biomacromolecular structure at atomic resolution is X-ray diffraction from a single crystal. A molecular structure solved to atomic resolution means that the position of each atom could be distinguished from those of all other atoms in three-dimensional space without applying additional assumptions concerning the structure of the molecule. X-rays are photons with wavelengths in the range of 1 to 1000 nm. Typical X-rays used in structure determination are K_α rays of Cu ($\lambda = 15.4$ nm) and Mo ($\lambda = 7.1$ nm). X-ray and neutron diffraction patterns can be detected when a wave is scattered by a periodic structure of atoms in an ordered array such as a crystal or a fiber (Blundell and Johnson, 1976; Rousseau, 1998; Woolfson, 1997).

Crystals are solids in which the molecules are arranged in a regular, symmetric and repeating fashion. The basic unit that describes a crystal is called the unit cell. A crystal is then constructed by translating the unit cells in three dimensions to fill a volume, the unit cell is constructed by translating the repeating motif (the lattice points) and the lattice is a regularly repeated arrangement of atoms. The unit cell of a lattice is the smallest and simplest unit from which the 3D periodic pattern can be produced. It may be made up from several identical asymmetric units in which case the lattice is generated by applying symmetry operators to the asymmetric units. Unit-cell dimensions are usually labeled by the letters a, b and c, while the dimensions of the molecule are usually labeled x, y and z. Lattice points are defined as the corners (vertices) of the unit cells. Lattice planes are a set of equivalent parallel planes constructed so that all lattice points lie on some member of the set. Planes passing through the opposite vertices of the unit cells cut the axes (a, b and c) precisely at the values corresponding to one unit translation of the lattice. However, increasingly finer-spaced lattice planes can be drawn. In three dimensions, lattice planes exist that cut one axis every a/h while cutting another at b/k and the third at c/l. These planes can be described by specifying the Miller indices, h, k and l. Parallel planes represented by three Miller indices (h, k, l) can be used to construct a reciprocal lattice. Different diffracted waves interfere with each other and produce an interference pattern. The diffraction patterns can be interpreted directly to give information about the size of the unit cell, information about the symmetry of the molecule and in the case of fibers, information about periodicity. The diffraction pattern has a reciprocal relationship with the object. A crystal is a 3D array of atoms in real space and its diffraction pattern is a 3D array in reciprocal space.

Diffraction refers to phenomena occurring when a light wave passes an object and is scattered in all directions. The scattered light in any particular direction is characterized by amplitude and phase terms. The pattern of the scattered waves is called the diffraction pattern of the object. X-ray photons (with wavelength, λ) are scattered by electrons in matter. For the scattered waves from a multi-electron system, each electron scatters the wave independently in all directions. But the scattered intensity (I_{sca}) is affected by the interference between the scattered waves depending on the phase difference ($2\pi\Delta/\lambda$, where Δ is the path difference for the two scattered beams between source and detector) between the waves for the two adjacent electrons according to

$$I_{sca} = 2(CI_o/r^2)(1 + \cos 2\pi\Delta/\lambda)$$

where I_o is the incident intensity, *r* is the distance from the scattering point to the detector and C is a proportionality factor that depends on the charge and mass of the electrons. The amplitude, A is expressed as

$$A^2 = 2(1 + \cos 2\pi\Delta/\lambda)$$

and is also dependent on the phase difference between the waves ($2\pi\Delta/\lambda$). If $\Delta = 0, \lambda, 2\lambda, \ldots, \cos 2\pi\Delta/\lambda = 1$, the interference scattering from two electrons is four times the scattering from the one electron ($I_{sca} = 2(CI_o/r^2)$), This is called total constructive interference or total reinforcement. If $\Delta = \lambda/2, 3\lambda/2, 5\lambda/2 \ldots, \cos 2\pi\Delta/\lambda = -1$ and the intensity of scattering from two electron is zero. This is total destructive interference. Thus the intensity of a scattered wave is determined by the ratio of path difference to wavelength. The scattering from an atom is characterized by the atomic scattering factor, *f*, which is the ratio of the amplitude scattered by the atom to that by a point electron. At zero scattering angle *f* is equal to the number of electrons in the atom.

For identical scatters located at each lattice point in a crystal, the total reinforcement occurs if the following Laue's equations are satisfied:

$$a(\cos\alpha - \cos\alpha_o) = h\lambda$$
$$b(\cos\beta - \cos\beta_o) = k\lambda$$
$$c(\cos\gamma - \cos\gamma_o) = l\lambda$$

The point scatters are equally spaced with distances a, b and c between adjacent pairs along *x*-, *y*- and *z*-axes. The angles, α_o, β_o and γ_o and α, β and γ are the angles that the incident and diffracted beams make with the three axes respectively. The integers, h, k and l are Miller indices. The diffraction pattern from a crystal will thus consist of spots of scattered intensity whose positions depend on the crystal lattice.

The Bragg equation provides a simpler physical relationship between the directions of incident and scattered rays by a crystal plane, i.e. a monochromic incident beam of wavelength λ will be reflected by a family of parallel crystal planes (h, k, l) if the incident angle is θ:

$$\theta = \sin^{-1} n\lambda/(2d)$$

where n is an integer and d/n can be viewed as the spacing between a family of planes with Miller indices (nh, nk, nl). To state it differently, if the incident rays with wavelength λ make an angle θ with a family of crystal planes, such that

$$2d\sin\theta = n\lambda$$

reinforcement of the scattered rays occurs in a direction that also makes an angle θ with the planes. The equation provides the relationship between the diffraction pattern and the lattice parameters. The intensity of the scattered ray is a function of $\sin\theta/\lambda$ because constructive interference occurs only at $\sin\theta/\lambda = n/2d$. In the Laue's equation, h, k and l, which are the Miller indices, define the integer number of wavelengths that result in an observed reflection from a three-dimensional crystal. Thus for a given set of Miller indices, Bragg law and Laue's equation are related

$$(h^2/a^2 + k^2/b^2 + l^2/c^2)^{1/2} = 2\sin\theta/\lambda = n/d$$

At a given Bragg angle, the intensity of rays diffracted by the lattice atoms is dependent on their atomic structure factors, *f*'s and their relative phases of diffracted rays. Therefore the determination of the complete structure of a molecule requires the phase information as well as the intensity and frequency information. The phase can be determined using the method of multiple isomorphous replacement where heavy metals or groups containing heavy element are incorporated into the diffracting crystals. The final coordinates of biomacromolecules are then deduced using knowledge about the primary structure and are refined by processes that include comparisons of calculated and observed diffraction patterns. Three-dimensional structures of proteins and their complexes (Blundell and Johnson, 1976), nucleic acids and viruses have been determined by X-ray diffraction and neutron diffraction (Roussseau, 1998). Protein Data Bank (http://www.rcsb.org/pdb) is the international publicly available repository for the 3D structures of biomacromolecules. The 3D structures of nucleic acids can be retrieved from Nucleic Acid Database (NDB) at http://ndbserver.rutgers.edu. The SWEET2 server of Glycosciences (http://www.glycosciences.de/modeling/sweet2/doc/index.php) provides tools for modeling the 3D structures of saccharides.

7.7.2 Crystallographic study of biomacromolecules

The following steps are typically used in the crystallographic study of biomacromolecules:

7.7.2.1 Preparation of biomacromolecular crystals. The most important factor in crystallizing a molecule is the purity of the sample. In most cases, a biomacromolecule must be better than 95% pure to produce a crystal. Two widely used strategies to facilitate crystallization, based on reducing the solubility of a biomacromolecule in a controlled manner, are vapor diffusion and microdialysis. The common vapor diffusion techniques are the hanging drop method (hanging the sample above the reservoir) and the sitting drop method (sitting the sample in a well surrounded by the reservoir). In the microdialysis, solvent is transferred by equilibrating the osmotic pressures of the sample and reservoir across a semipermeable membrane.

7.7.2.2 Preparation of isomorphous heavy-atom derivatives. The multiple isomorphous replacement technique has been commonly used to solve biomacromolecular structures. It requires the parent crystal and at least two heavy-atom derivatized crystals identical in space group and molecular structure. The common technique uses reagents containing heavy atoms and allows them to diffuse into the crystal.

7.7.2.3 Determination of X-ray diffraction patterns for parent and isomorpous crystals. The spacing of the indexed reflections is inversely proportional to the lengths of the crystal unit cell. Since the relative spacing between reflections is needed for determining the lengths of the unit cell axes, it is important to obtain an undistorted diffraction pattern. Each reflection in the diffraction pattern can be assigned to a unique set of Miller indices from the Laue conditions for diffraction. The pattern is indexed in term of the Miller indices.

7.7.2.4 Determination of structure factors, location of heavy atoms for isomorpous crystals. The electron density distribution of the heavy-atom isomorphous crystal is the sum of the electron densities of the parent crystal and of the heavy-atom substituents, i.e.:

$$F_{PH}(h, k, l) = F_P(h, k, l) + F_H(h, k, l)$$

where F_{PH}, F_P and F_H are structure factors of the heavy-atom isomorphous derivatives, parent crystal and heavy-atoms. To locate the heavy atoms in the macromolecular crystal, the Patterson map is used. It is possible to calculate a difference isomorphous Patterson map between them using the measured structure factor amplitudes, $|F_{PH}(h, k, l)|$ and $|F_P(h, k, l)|$ if both crystals (heavy-atom isomorphous and parent) are available and the presence of the heavy atom in the isomorphous derivative causes a change in scattering intensity sufficient to yield measurable differences between the two structure factors. Once the heavy atom is found, F_H can be computed. The phase of F_P is selected to evaluate F_{PH} and F_P using the known F_H. This process is repeated until self consistency is reached.

7.7.2.5 Estimation of phases of the structure factor. Both the phase and amplitude of the contribution of the heavy atom to the structure factor are computed to yield F_H.

7.7.2.6 Least square refinement of a structural model. A least square refinement is applied to minimize the difference in the adjustable parameters used in the calculation of an X-ray scattering pattern between observed structure factor and those calculated from particular model or technique.

7.7.2.7 Calculation of an electron density map. Several heavy-atom derivatives are prepared and isomorphous replacement has been used to estimate phases for all $F_P(h, k, l)$. These phases are used to calculate an electron density map of the crystal and all data to a certain resolution.

7.7.2.8 Construction of a molecular model from the electron density map. At 6 Å resolution, the macromolecule usually appears as a blob of electron density and the chain backbone is generally unrecognizable. At 3.0 Å resolution, it is possible to trace the path of the macromolecular chain backbone. The double helices of nucleic acids are traced readily. At 2.0 Å resolution, almost all protein side chains, nucleic acid nucleotides and polysaccharide glycoses are visible. A model of the protein, nucleic acid or glycan can be constructed if the amino acid, nucleotide or monosaccharide sequence is known. At higher resolution, individual atoms begin to be seen. It is possible to identify amino acid side chains, nucleotides and glycose units directly from the electron density map.

7.7.2.9 Refinement of the structure. The accurate structural model can be built from a higher resolution of a crystal structure. An improvement in the electron density map resulting from additional refinement yields higher resolution image. Increasingly, the computer is used to elucidate/refine structures of biomolecules (Pretsch *et al.*, 2003).

Two techniques currently provide 3D structural information of biomacromolecules. An X-ray crystal structure is an atomic model based on the interpretation of a high resolution image, and an NMR structure is an atomic model based on the interpretation of a set of interatomic distances. In contrast to NMR, which provides identifiable interatomic distances, the maps of interatomic vectors provided by X-ray data are essentially impossible to deconvolute without an intermediate step, called phase determination. Phase determination constitutes the imaging step and completes a close underlying connection between crystallography and microscopy. Thus, in a sense, crystallography is more complete with respect to the quality of information it provides. NMR provides crucial evidence regarding behavior in solution. X-ray diffraction provides, via refinement techniques, the most accurate atomic models available. Some guidelines, as presented in Table 7.11, are essential for prospecting the structural database.

A procedure called ‘cross-validation’ (Brunger, 1992) provides a quantitative estimate for the likely errors in a refined model by withholding some of the data from the

TABLE 7.11 Reliability characteristics of atomic coordinate files

Characteristics		Well-defined features
Resolution limits	6.0–4.5 Å	Placement of secondary structures
	3.0 Å	Chain tracing
	2.5 Å	Side-chain orientation, isotropic thermal parameters
	2.0 Å	Side-chain, bound water identification
	1.8 Å	Alternate side chain orientation
	1.5 Å	Anisotropic thermal parameters
	1.2 Å	Hydrogen atoms
Source of phases	Multiple isomorphous replacement	Free of model bias; but noisy due to lack of isomorphism
	Multi-wavelength anomalous diffraction	In general the most reliable source of phases; isomorphism is nearly perfect
	Molecular replacement	Widely used, errors due to model bias are variable and difficult to detect, correct
Model-independent map refinement	Noncrystallographic symmetry (NCS)	Widely observed and enforced.
	Solvent flatness (SF)	Strength proportional to solvent volume-fraction
	Histogram matching (HM)	Complementary to NCS and SF
Refinement	$R_{free} <\sim 1.2\,m \times R_{cryst}$	Cross-validation statistic; indispensable to avoid model bias due to overfitting
Stereochemistry	Atypical ($\phi\psi$) angles	Less than 1–2% of residues
	Bond length	±0.005 Å
	Bond angle	±2°

refinement process. A random sample of the X-ray data (~1000 reflections) is reserved as a reference or test set and these reflections are not used to constrain the model. The crystallographic R-factor for the test set forms the 'free R-factor' (R_{free}), which should improve in parallel with the R-factor for the data being used to constrain the model. Structures for which R_{free} is more than 20% higher than the quoted R-factor almost certainly contain errors or artifacts due to over-fitting of the experimental data.

Few structure reported at resolutions lower than about 3.5 Å are free of ambiguity regarding side chain placements or identity. Structures reported at resolutions between 2.9 Å and 3.5 Å probably have the correct path for the main chain, but require considerable guess work in the placement of side chain groups. At 2.5 Å resolution, maps begin to reveal correct side chain orientations. At 2.0 Å, we can often sequence much of the protein from the side chain identification in the map. At 1.5 Å resolution, five-membered rings have well-resolved holes and alternate conformations can be confidently identified for side chains. Beyond about 1.2 Å, we can begin to locate hydrogen atoms in electron density maps.

Interactive computer graphic tools extend the resources for learning about protein structure far beyond what can be conveyed in even the most carefully prepared static images. Atomic coordinates are available for most three-dimensional structures and can be accessed from the World Wide Web (PDB). Two different computer programs are widely used that no student should be without. KineMage (http://kinemage.biochem.duke.edu/) incorporates interactive graphic illustration kinemages into publications by Protein Science. RasMol (http://www.umass.edu/microbio/rasmol/) is a complementary program that offers interactive control over the display format.

Proteins exhibit regularities at several levels, which can be identified only by the fact that they recur approximately, but frequently, in different contexts. Much insight into the structures of proteins is therefore statistical in nature, coming from the comparison of large numbers of different proteins.

7.8 REFERENCES

Bax, A. (1989) *Annual Reviews in Biochemistry*, **58**, 223–56.

Bell, J.E. (ed.) (1981) *Spectroscopy in Biochemistry*, CRC Press, Boca, Raton, FL.

Blundell, T.L. and Johnson, L.N. (1976) *Protein Crystallography*, Academic Press, New York.

Bock, K. and Pedersen, C. (1983) *Advances in Carbohydrate Chemistry and Biochemistry*, **41**, 27–66.

Bock, K., Pedersen, C. and Pedersen, H. (1984) *Advances in Carbohydrate Chemistry and Biochemistry*, **42**, 193–225.

Brunger, A. (1992) *Nature*, **355**, 472–4.

Campbell, I.D. and Dwek, R.A. (1984) *Biological Spectroscopy*, Benjamin/Cummings, Menlo Park, CA.

Donovan, J.W. (1969) *Journal of Biological Chemistry*, **244**, 1961–9.

Evans, J.N.S. (1995) *Biomolecular NMR spectroscopy*, Oxford University Press, London.

Fasman, G.D. (ed.) (1996) *Circular Dichroism and the Conformational Analysis of Biomolecules*, Plenum Press, New York.

Gorin, P.A.J. (1980) *Advances in Carbohydrate Chemistry and Biochemistry*, **38**, 13–104.

Greenfiel, N. and Fasman, G.D. (1969) *Biochemistry*, **8**, 4108.

Greve, J., Puppels, G.J. and Otto, C. (1999) *Spectroscopy of Biological Molecules*: *New Directions*, Kluwer, Boston, MA.

Hammes, G.G. (2005) *Spectroscopy for the Biological Sciences*, John Wiley & Sons, Hoboken, NJ.

Jimenez-Barbero, J. and Peters, T. (eds) (2002) *NMR Spectroscopy of Glycoconjugates*, John Wiley. New York.

Karplus, M. (1959) *Journal of Chemical Physiology*, **30**, 11–5.

Lakowicz, J.R. (1999) *Principles of Fluorescence Spectroscopy*, 2nd edn, Kluwer Academic/Plenum, New York.

Mantsch, H.H. and Chapman, D. (eds) (1996) *Infrared Spectroscopy of Biomolecules*, Wiley-Liss, New York.

Moore, P.B. (1995) *Acc. Chemistry Research*, **28**, 251–6.

Parker, F.S. (1983) *Application of Infra-red, Raman and Resonance Raman Spectroscopy in Biochemistry*, Plenum Press, New York.

Pohl, F.M. and Jovin, T.M. (1972) *Journal of Molecular Biology*, **67**, 3759–6.

Pretsch, E., Tóth, G., Munk, M.E. and Badertscher, M. (2003) *Computer-Aided Structure Elucidation.* John Wiley & Sons, New York.

Roberts, G.C.K. (1993) *NMR of Macromolecules*. IRL Press, Oxford.

Roussseau, J. (1998) *Basic Crystallography*, John Wiley & Sons, New York.

Singh, B.R. (2000) *Infrared Analysis of Peptides and Proteins*, Oxford University Press, Oxford.

Stassinopoulou, C.I. (1994) *NMR of Biological Macromolecules*, Springer-Verlag, New York.

Thomas, G.J. (1969) *Biopolymers*, **7**, 325.

Timasheff, S.N. and Gorbunoff, M.J. (1967) *Annual Reviews in Biochemistry*, **36**, 13.

Tugarinov, V., Hwang, P.M. and Kay, L.E. (2004) *Annual Reviews in Biochemistry*, **73**, 107–46.

Varani, G. and Tinoco, I. Jr. (1991) *Quarterly Review in Biophysics*, **24**, 479–532.

Wagner, G., Neuhaus, D., Wörgötter E. *et al.* (1986) *Journal of Molecular Biology*, **187**, 131–5.

Wishart, D.S. and Nip, A.M. (1998) *Biochemical Cell Biology*, **76**, 153–63.

Woolfson, M.M. (1997) *An Introduction to X-ray Crystallography*, 2nd edn, Cambridge University Press, Cambridge, UK.

Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, John Wiley & Sons, New York.

World Wide Webs cited

BioMagResBank:	http://www.bmrb.wisc.edu/
DICHROWEB:	http://public-1.cryst.bbk.ac.uk/cdweb/
KineMage:	http://kinemage.biochem.duke.edu/
NMR DB of Glycosciences:	http://glycosciences.de/sweetbase/nmr/
Nucleic Acid Database:	http://ndbserver.rutgers.edu.
Protein Data Bank:	http://www.rcsb.org/pdb
RasMol:	http://www.umass.edu/microbio/rasmol/
SDBS:	http://www.sist.go.jp/RIODB/SDBS/menu-e.html
SWEET2:	http://www.glycosciences.de/modeling/sweet2/doc/index.php

CHAPTER 8

STUDIES OF BIOMACROMOLECULAR STRUCTURES: CHEMICAL SYNTHESIS

8.1 RATIONALE

Nucleic acids, especially DNA can be prepared easily by biochemical techniques such as cloning and polymerase chain reaction (Innis *et al*., 1999; Mullis *et al*., 1994), and protein sequences that are encoded by DNA can therefore be produced and manipulated through recombinant DNA technology (Greene and Rao, 1998). There is no information carrier that encodes a particular glycan and so biochemical synthesis of glycan with processes akin to nucleic acids and proteins is not available. Enzymatic approaches to peptide and saccharide syntheses have been attempted (Kullmann, 1987; Sears and Wong, 2001). In addition, syntheses of oligo- and polynucleotides/oligo- and polypeptides have become efficient with the advent of automated solid-phase synthesis (SPS) and multiple (combinatorial) synthetic techniques (Fenniri, 2000; Kates and Albericio, 2000) of which knowledge on chemical syntheses of these bio-oligomers and polymers are essential.

Chemical synthesis of pharmacologically active oligonucleotides (Hardewijn, 2005), oligopeptides (Pennington and Dunn, 1994; Howl, 2005) and oligosaccharides (Ernst *et al*., 2000) are often preferred for economic considerations, convenience and quantities. Their analogues can be made using combinatorial technique and as substrates/inhibitors in studies of enzymology and pharmacology. Investigations into their antigenicity and uses in the epitope mapping provide useful information directed toward diagnostic targets and vaccine development. Chemically synthesized sequence defined oligonucleotides are required for genome syntheses and various aspects of genomic research. Sequence specific oligomers are chemically synthesized *in silico* for microarray analyses. Tailor-made model oligomers serve to study conformational behavior using a variety of physicochemical methods. Therefore chemical syntheses of oligomeric and polymeric nucleotides/peptides/saccharides are valuable tools in investigations of the structures and applications of biomacromolecules.

8.2 SYNTHETIC STRATEGY: CONVENTIONAL APPROACH

The conventional approach to the synthesis of biomacromolecules usually starts with the constituent monomers, i.e. respective nucleotides, α-amino acids and monosaccharides or

---

Biomacromolecules, by C. Stan Tsai
Copyright © 2007 John Wiley & Sons, Inc.

their derivatives. Biomacromolecules are sensitive to a wide range of chemical reactions and therefore only the mildest reaction conditions can be used in the assembly of biopolymeric chains. The key to their synthesis is the specific and sequential formation of the respective phosphodiesteric, amide and glycosidic linkages. Essentially these linkages are formed via a dehydration reaction in three steps:

1. **P***rotection*: protection of functional groups that do not participate (or may be destroyed) in the subsequent coupling reaction. The protecting groups must be easily removed upon completion of the synthesis without detrimental effect to the assembled chain.

   $\text{X-R}_m\text{–OH} + \text{Y-A} \rightarrow \text{Y–X-R}_m\text{–OH} + \text{HA}$ where $\text{X} = \text{—NH}_2$ or —OH and Y = protecting group

2. *Coupling*: formation of the biopolymeric linkage that joins two monomers or the new monomer to the chain by either activation-condensation or direct dehydration coupling.

   $$\text{Y–X-R}_m\text{–OH} + \text{H-R}_n\text{-X}'\text{–Y}' \rightarrow \text{Y–X-R}_m\text{–R}_n\text{-X}'\text{–Y}' + \text{H}_2\text{O}$$

3. *Deprotection*: removal of protecting groups via mild treatment to give the final product.

   $$\text{Y–X-R}_m\text{–R}_n\text{-X}'\text{–Y}' \rightarrow \text{X-R}_m\text{–R}_n\text{-X}'$$

8.2.1 Protection and deprotection of common functional groups

Two functional groups, amino and hydroxy, are commonly encountered in the chemical synthesis of biomacromolecules that need to be protected. These groups must be cleaved at certain stages of synthetic processes selectively under conditions that do not interfere with the stability of the biopolymeric bonds assembled. Furthermore, racemization should not occur during protection and deprotection operations. Because of the variety and multiplicity of the functional groups that need to be protected, protecting group orthogonality is a criterion that is especially important for the success of biomacromolecular syntheses. Orthogonality means that subsets of protecting groups present in a molecule can be cleaved selectively under certain reaction conditions, while all other protecting groups remain intact. Tables 8.1 and 8.2 list the commonly used protecting groups and their removal (cleavage) for the amino and hydroxy groups respectively. In principle, an amino group can be blocked reversibly by acylation, alkylation and alkyl-acylation. Protecting groups can be then cleaved by acid hydrolysis, base cleavage, reduction/oxidation, nucleophilic substitution and photolysis.

Orthogonal protection is an important consideration for glycoside synthesis involving polyhydroxyl groups. The types of protection reactions that can be carried out at hydroxyl groups of glycoses include esterification and etherification. Acetylation is commonly employed during the purification of glycoses/glycans and as such is carried forward in glycoside synthesis. Tosylation is used to protect secondary hydroxyl groups without blocking the glycosidic (C1) hydroxy. Furthermore, the tosyl (Ts) group can be easily displaced with other nucleophiles and offers a convenient route to the synthesis of glycose/glycan derivatives. It is noted that the primary hydroxyl group is a more effective nucleophile and less hindered than the secondary hydroxy groups. Triphenylmethyl (trityl, Tr) group and its derivatives, such as 4-methoxytriphenylmethyl (monomethoxytrityl, MMTr) and 4,4′-dimethoxytriphenylmethyl (dimethoxytrityl, DMTr) are commonly used selectively to protect primary hydroxy of glycoses and pentose moieties of nucleotides,

TABLE 8.1 Selected amino protecting groups, Y in Y—HN · R

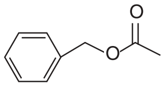
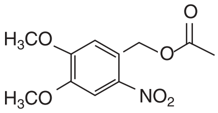
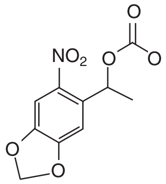
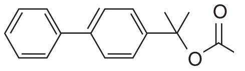
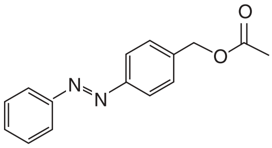
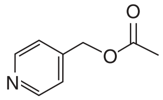
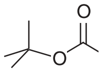
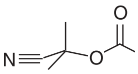
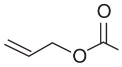
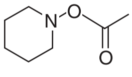
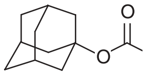
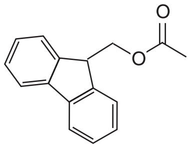
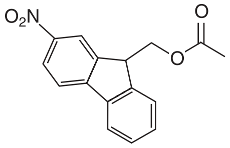
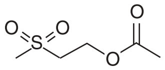
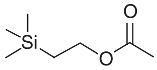
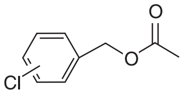
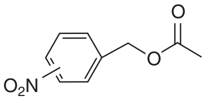
Group (Y)	Symbol	Formula for Y	Cleavage condition
Benzyloxycarbonyl	Z (Cbz)		HBr/AcOH, TFA, Na/liqNH_3, H_2/Pd
6-Nitroveratryloxycarbonyl (4,5-dimethoxy-2-nitrobenzoxycarbonyl)	Nvoc		As Z and photolysis
Methylnitropiperonyloxy carbonyl	MNPOC		Photolysis
2-(Biphenyl-4-yl)-2-propoxycarbonyl	Bpoc		80% AcOH, $MgClO_4$
4-(Phenyldiazenyl)-benzyloxycarbonyl	Pz		HBr/AcOH, Na/liqNH_3, H_2/Pd
Isonicotinyloxycarbonyl	iNoc		Zn/AcOH, H_2/Pd, acid stable
tert-Butoxycarbonyl	Boc (tBoc)		TFA, TFA/CH_2Cl_2, HCl/organic solvent
2-Cyano-*tert*-butoxycarbonyl	Cyoc		Weak base (aq. KCO_3, triethylamine)
Alloxycarbonyl	Aloc		Pd(0)/nucleophile
Piperidinyloxycarbonyl	Pipoc		Electrolysis, H_2/Pd
Adamantyl-1oxycarbonyl	Adoc		TFA
Fluorenyl-9-methoxycarbonyl	Fmoc		Piperidine, 2-aminoethanol, morpholine, liq NH_3

TABLE 8.1 ***continued***

Group (Y)	Symbol	Formula for Y	Cleavage condition
(2-Nitrofluoren-9-yl)methoxycarbonyl	NO_2Fmoc		Photolysis
Methylsulfonyl-ethoxycarbonyl	Msc		Based catalyzed β-elimination
2-(Trimethylsilyl)-ethoxycarbonyl	Teoc		F$^-$, e.g. TFA, tetrabutylammonium fluoride

Notes: 1. Various Cbz (Z) derivatives, such as chlorobenzyloxycarbonyl (ClZ) and nitrobenzyloxycarbonyl (NO_2Z), which are more stable to HBr/AcOH but more labile to H_2/Pd have been employed:

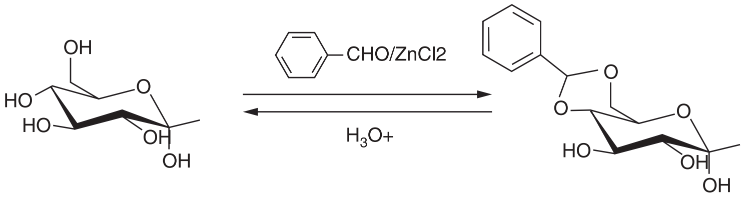




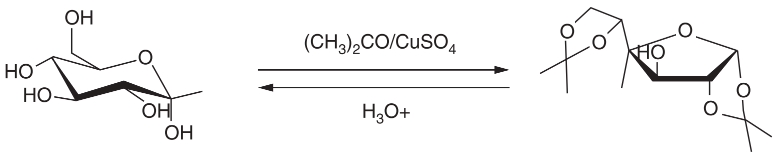
Ortho-nitrobenzyloxycarbonyl group is used as a photolabile linker for the attachment of encoding molecules in combinatorial synthesis of peptides.

2. For peptide synthesis, amino-protecting groups with different labilities toward acidic deblocking (cleaving) agents are preferred over base cleavage, because most peptides are sufficiently stable under moderately acidic conditions. These protecting groups include Cbz or Z, Boc, Fmoc, Aloc and their derivatives

because of the regioselectivity of the bulk trityl group for the primary hydroxy compared to the secondary hydroxyl groups. The acid labilities of trityl derivatives increase as the number of methoxy groups increase. Carbonyl compounds such as acetone and benzaldehyde are used to protect the *vicinal*-diol (glycol) functionality of glycoses. The treatment of glycoses with benzaldehyde forms 4,6-*O*-benzylidene-D-glycopyranose:



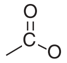
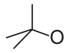
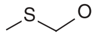
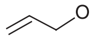
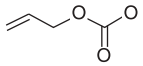
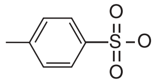
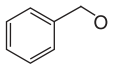
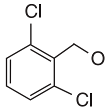
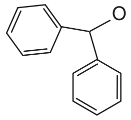
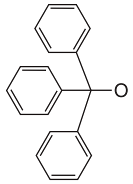
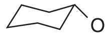
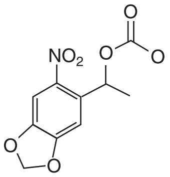
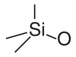
while acetonation yields 1,2:3,4-di-*O*-isopropylidene-D-glycofuranose:



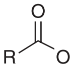
8.2.2 Protection and deprotection specific to peptide synthesis

Peptide synthesis becomes further complicated by the presence of functional groups in the side chains, other than hydroxyl and amino groups that need to be selectively protected.

TABLE 8.2 Hydroxy protecting groups, Y—O · R

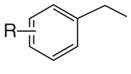
Group (Y)	Symbol	Formula for Y—O	Cleavage condition
Acetyl	Ac		CH_3ONa/CH_3OH
tert-Butyl	tBu		Trifluoroacetic acid (TFA), HCl/TFA, conc HCl at 0°C, 10 min
Methylthiomethyl	Mtm		CH_3I in wet acetone ($NaHCO_3$)
Allyl	Al		Pd^0, nucleophile, e.g. *tert*BuO^-
Allyloxycarbonyl	Aloc		Pd^0, nucleophile, e.g. *tert*BuO^-
Toluenesulfonyl (tosyl)	Ts		Base
Benzyl	Bzl, Bz		HF, HBr/dioxan, Na/liq NH_3, H_2/Pd
2,6-Dichlorobenzyl	Dcb		Strong acids, stable to TFA
Diphenylmethyl (benzhydryl)	Dpm		TFA (reflux), H_2/Pd
Triphenylmethyl (trityl)	Tr		p-Toluenesulfonic acid, TFA, H_2/Pd
Cyclohexyl	Cy		Trifluoromethane sulfonic acid with thioanisole
Methylnitropiperonyloxy carbonyl	MNPOC		Photolysis
Trimethylsilyl	Tms		NH_4OH, Bu_4NF

Notes: 1. Acylation/deacylation is commonly used for purification/analysis/preparation of glycoses and glycans. Acyl groups used are:



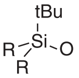
where R = CH_3—, $ClCH_2$—, Cl_3CCH_2—, $(CH_3)_3C$—, $CH_3COCH_2CH_2$—, C_6H_5—, and ClC_6H_4—,

2. Various benzyl derivatives are used to protect polyhydroxy groups of glycoses and their removal conditions (in parentheses) are:



where R = H (Pd), OCH_3 (TFA), $OCOCH_3$ ($NaOCCH_3$), Cl/Br (Pd, secondary amine).

3. A number of tertiary silyl derivatives, besides TMS, such as *tert*-butyldimethylsilyl (TBDMS), and *tert*-butyldiphenylsilyl (TBDPS) have been used as protecting polyhydroxyl groups in glycan synthesis:



where R = CH_3 (TBDMS); C_6H_5 (TBDPS).

4. In the phosphotriester synthesis of oligonucleotides, the hydroxy of nucleotide phosphate is protected with aryl group such as 2-/4-chlorophenyl (CP) or 2,5-dichlorophenyl (DCP) group. MNPOC is used as the photolabile protecting group for *in silico* photolithographic oligonucleotide synthesis (DNA chip).

Because of the different requirements with respect to selectivity, a distinction must be made between the intermediary (temporary) and semipermanent protecting groups. The intermediary groups are used for temporary protection of the amino or carboxyl function involved in subsequent bond formations. These groups must be cleaved selectively at each step prior to peptide bond formation. Thus orthogonality is especially important for selecting the intermediary protecting groups in a peptide synthesis. The semipermanent protecting groups are usually cleaved only at the end of the peptide synthesis. The protection/deprotection chemistry for amino and hydroxy group has been dealt in the preceding section 8.1.1. Other functional groups specific to peptide synthesis are summarized in Tables 8.3, 8.4 and 8.5.

The guanidino group of Arg is usually protected by protonation under normal reaction conditions due to its strongly basic character. The indole ring of Trp usually does not require any protecting measures. The broad variety of synthetic conditions permits the incorporation of unprotected Trp. The methionine thioether group usually does not cause severe complications during peptide synthesis.

8.2.3 Coupling reaction

Phosphodiesteric/peptidic/glycosidic bond formation (coupling reaction) is a nucleophilic substitution reaction of a hydroxy/amino/hydroxy group at a phosphoesteric/carboxyl/acetal group, resulting in dehydration. The coupling reaction can be conducted in two procedures; activation-substitution (via stable activated intermediate) and direct condensation (via transient reactive intermediate).

8.2.3.1 Activation-substitution. To promote coupling of biomonomers via a nucleophilic substitution, which has to be performed under mild conditions, the electrophilic (esteric/carboxylic/acetal) site is activated to increase its electrophilicity. This is achieved by an introduction of electron-withdrawing moieties (decrease the electron density at the electrophilic site), thereby favoring the subsequent nucleophilic attack. Some of the common activated structures are illustrated in Table 8.6.

8.2.3.2 Direct condensation. The direct condensation of biomonomers between the two coupling centers can be initiated by the use of coupling reagents such as carbodiimides:

$$-X-OH+H-y-\xrightarrow{\text{R–N-C=N–R}'}-X-Y-+\text{R-HN-C(=O)-NHR}'$$

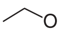
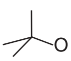
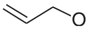
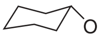
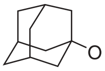
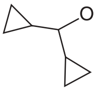
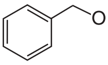
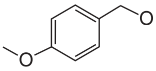
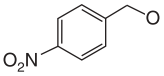
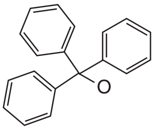
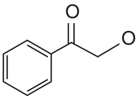
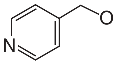
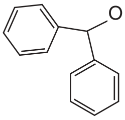
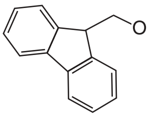
where X represent phosphate of nucleotide, carboxyl of amino acid, or acetal of glycose component and Y is hydroxyl of nucleotide, amino of amino acid or hydroxyl of glycose moiety (Table 8.7). Enzymes (e.g. transferases) are also used as coupling agents in biooligomer synthesis.

8.3 SYNTHETIC STRATEGY: SOLID PHASE APPROACH

8.3.1 General concept

The concept of peptide synthesis on a solid support known as solid phase peptide synthesis was developed in 1963 (Merrifield, 1963) and the concept has been extended and generalized to organic synthesis on polymeric supports. The method is particularly attractive to the synthesis of biomacromolecules (Bayer, 1991; Fields, 1997; Kates and Albericio, 2000; Merrifield 1985).

TABLE 8.3 Selected carboxyl protecting groups, Y in Y—C(=O) · R

Group (Y)	Symbol	Formula for Y	Cleavage condition
Methyl	Me	—O	Alkaline or enzymatic hydrolysis
Ethyl	Et		Alkaline or enzymatic hydrolysis
tert-Butyl	tBu		TFA, HBr/AcOH, HCl/AcOH, $BF_3 \cdot OEt_2$/AcOH
Allyl	Al		Pd(0)/nucleophile
Cyclohexyl	Cy		Stong acid
1-Adamantyl	1-Ada		TFA
Dicyclopropylmethyl	Dcpm		1% TFA/CH_2Cl_2
Benzyl	Bzl		Alkaline hydrolysis, HBr/AcOH, liq HF, H_2/Pd
4-Methoxybenzyl	Mob		As Bzl, HCl/nitromethane, TFA/anisole
4-Nitrobenzyl	Nbz		As Bzl, but stable in HBr/AcOH
Triphenylmethyl (Trityl)	Trt		1% TFA/CH_2Cl_2
Phenacyl	Pac		Sodium thiophenolate, H_2/Pd, Zn/AcOH
Pyridyl-4-methyl (4-picolyl)	Pic		Alkaline hydrolysis, electrolytic reduction, H_2/Pd
Diphenylmethyl (Benzhydryl)	Dpm		H_2/Pd, TFA at 0°C, HCl/AcOH, $BF_3 \cdot OEt_2$/AcOH at 25°C
9-Fluorenylmethyl	Fm		Tetrabutylammonium fluorIde, piperidine,

In solid phase synthesis (SPS) of biopolymers, the peptide/nucleotide/glycoside chain is assembled in the usual manner (conventional approach) from the C-/3′-/NR-end. The first monomer unit of the biopolymer to be synthesized is connected via its carboxyl/hydroxyl/hydroxyl group to an insoluble polymer. A necessary prerequisite is that anchoring groups (linkers) are introduced into the polymeric material (step 1). The first protected monomer is then reacted with the functional group of the linker (step 2). The

TABLE 8.4 Selected thiol protecting groups, Y in Y—S · R

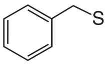
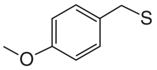
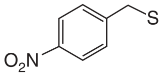
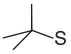
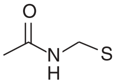
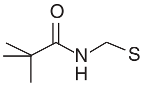
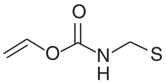
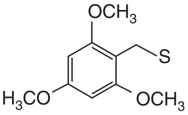
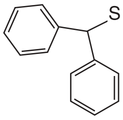
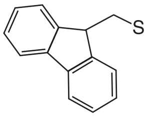
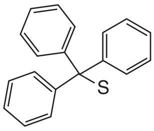
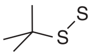
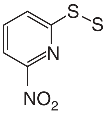
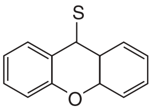
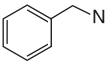
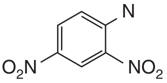
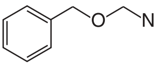
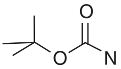
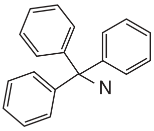
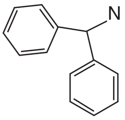
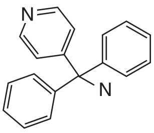
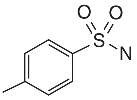
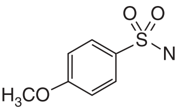
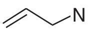
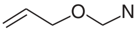
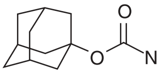
Group (Y)	Symbol	Formula for Y—S	Cleavage condition
Benzyl	Bzl		Na/liq NH_3
4-Methoxybenzyl	Mob		HF, TFA, trifluoromethane sulfonic acid, Hg^{II}
4-Nitrobenzyl	Nbz		
tert-Butyl	tBu		2-Nitrophenylsulfenyl chloride
Acetamidomethyl	Acm		Hg^{II}, I_2, 2-nitrophenylsulfenyl chloride followed by reduction
Trimethylacetamidomethyl	Tacm		Hg^{II}, I_2, $AgBF_4$
Allyloxycarbonyl-aminomethyl	Alocam		Pd^0/Bu_3SnH
2,4,6-Trimethoxybenzyl	Tmb		TFA
Diphenylmethyl	Dpm		TFA, HF, HBr/AcOH, 2-nitrophenylsulfenyl chloride then reduction
9-Fluorenylmethyl	Fm		Piperidine
Triphenyl (trityl)	Trt		HF, TFA, I_2, Hg^{II}, $AgNO_3$, 2-nitrophenylsulfenyl chloride then reduction
tert-Butylsulfenyl	StBu		
	Npys		
	Xan		

TABLE 8.5 Imidazole protecting groups, Y in Y—Nim R

Group	Symbol	Formula for Y—N^{im}	Cleavage condition
Benzyl	Bzl		Na/liq. NH_3
2,4-Dinitrophenyl	Dnp		2-mercaptoethanol, thiophenol
Benzyloxymethyl	Bom		Strong acids
tert-Butoxycarbonyl	Boc		TFA, HBr/AcOH, HF
Triphenylmethyl	Trt		TFA
Diphenylmethyl	Dpm		TFA (1 h), 6N HBr/AcOH (3 h)
Pyridyldiphenylmethyl	Pdpm		Catalytic or electrolytic reduction
4-toluenesulfonyl	Tos		Strong acids
4-methoxybenzenesulfonyl	Mbs		TFA/$(CH_3)_2S$
Allyl	Al		Pd reduction, nucleophile
Allyoxymethyl	Alom		Pd reduction, nucleophile
1-Adamantyloxycarbonyl	Adoc		

temporary protecting group is removed (step 3) and the next monomer unit is coupled (step 4). Steps 3 and 4 are then repeated (step 5) until the required peptide/nucleotide/glycoside sequence has been assembled. Finally, the covalent bond between the linker moiety and the biopolymer chain is cleaved (step 6). In many cases the semipermanent side-chain protecting groups may be simultaneously removed. When syntheses are carried out on a polymeric support there is no need to perform the tedious and time-consuming isolation and purification of all intermediates, as would be necessary in conventional synthesis. The product of all the reactions (the growing biopolymer chain) remains bound to the support during steps 3 to 5, and excess reagents and by-products are removed by fil-

TABLE 8.6 Activation of electrophilic site for coupling reaction

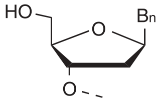
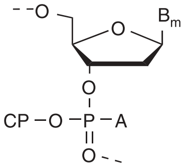
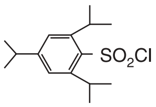
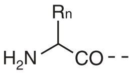
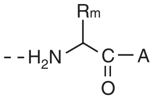
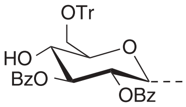
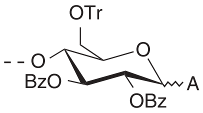
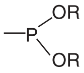
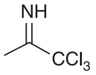
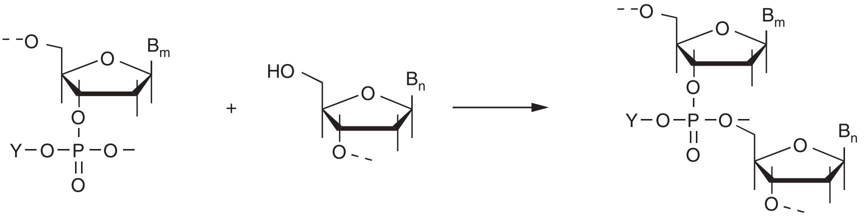
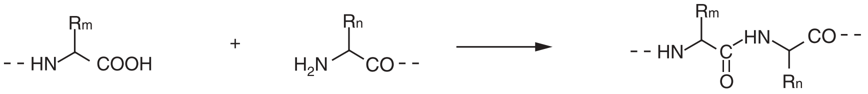
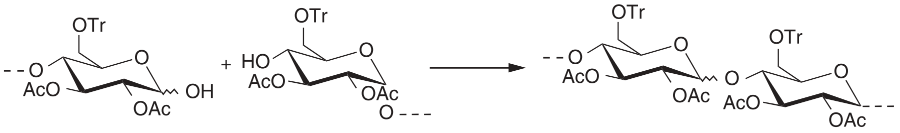
Coupling partners		Activator Group, A	Activator Reagent
Nucleophilic partner	Activated electrophilic partner		
Nucleotidic synthesis:			
		—Cl	
Peptidic synthesis			
		—Cl	CO_2Cl, $POCl_5$
		—N_3	NH_2NH_2/HNO_2
		—OCOR	RCOCl
Glycosidic synthesis:			
		—Cl, Br, —F	e.g. $SnCl_2/AgClO_4$
			PCl_3
			Cl_3CH_2CN/Base
		—SR —SOC_6H_5	RS^-

TABLE 8.7 Direct condensation for biomolecular synthesis

Bond formation	Condensation reaction
Nucleotidic	
Peptidic	
Glycosidic	

Note: The wavy glycosidic bond represents either α- or β-anomeric configuration.

tration/elution. Chemical database for SPS can be found at http://www.accelrys.com/products/chem_databases/databases/solid_phase_synthesis.html

SPS does suffer from several limitations:

- The final product of a synthesis carried out on a polymer support is only a homogeneous compound, if all deprotection and coupling steps proceed quantitatively.
- A large excess of each biomonomer component is required in the corresponding coupling reaction in order to achieve complete conversion.
- There is a risk of undesirable side reactions during activation, coupling and deprotection.
- Swelling properties of the polymeric resin and diffusion of the reagents are important parameters for the success of a SPS.
- The final product may be damaged, if drastic conditions are required to cleave the biopolymer chain from the polymer.

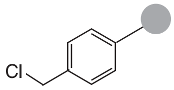
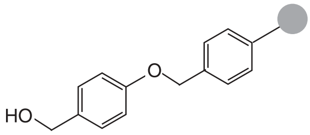
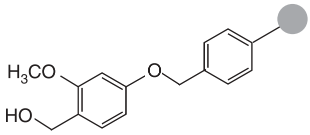
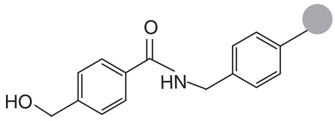
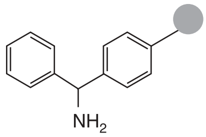
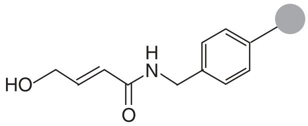
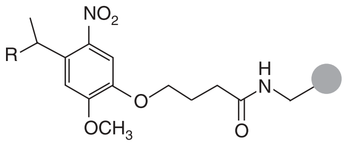
8.3.2 Solid-phase polymer support

The polymer support must inevitably be chemically inert, mechanically stable, completely insoluble in the solvents used and easily separated by filtration/elution. It must contain a sufficient number of reactive sites where the first biomonomer component can be attached. Interaction between the biopolymer chains bound to resin should be minimal. Table 8.8 gives some of the polymeric resins commonly used in the peptide synthesis. The introduction of anchoring groups (linkers) on a polymeric support is the precondition for application of solid phase in biopolymer synthesis and organic syntheses (James, 1999). Handles in the form of bifunctional linker moieties may be attached to the polymer matrix. One of functional groups fulfils the requirements for a protecting group allowing for cleavage under mild conditions, while the other is used for attachment to the resin. Correct selection of the handle allows yield optimization with respect to the attachment and cleavage steps.

Initially, a copolymer of polystyrene with 1–2% divinylbenzene as cross-linker was used in peptide SPS. The dry resin beads are normally 20–80 µmm in diameter and are able to swell to two- to six-fold volume in the different organic solvents used for peptide synthesis (e.g. acetonitrile, *N*,*N*-dimethylformamide, dioxan, dichloromethane and tetrahydrofuran). Consequently, the polymer support suspended in these solvents is a well-solvated gel with mobile polymeric chains. It was shown that the peptides are distributed uniformly within the gel matrix. Approximately 10^{12} polypeptide chains can locate on one polystyrene/divinylbenzene bead of 50 µm diameter, and with a resin loading of 0.3 mmole peptide per gram resin. The heterogeneous reactions used in SPS are usually two- or three-fold slower than homogeneous reactions. The optimal loading with the first amino acid ranges from 0.2 to 0.5 mmol per gram resin.

Cross-linked poly(dimethylacrylamide) was introduced as a more hydrophilic support. This forms a gel and is highly solvated by solvents appropriate for biopolymer synthesis. These polymers are stable to moderate pressure and may be used in columns for continuous-flow SPS. A hydrophilic polymer called TentaGel is obtained by grafting polyethylene glycol (PEG) chains on to polystyrene beads (polystrene core with polyethylene glycol arms). TentaGel (although less robust than polystrene physicomechanically) has desirable characteristics for SPS because the attached reacting groups project out in solution, providing for better reactivity. TentaGel has been used in SPS of glycans (St Hilaire and Meldal, 2000). Other support matrices for SPS of biomacromole-

TABLE 8.8 Some resins with linkers used in SPS

Resin	Linker structure	Remark
Chloromethyl (Merrifield) resin		Attachment by nucleophilic substitution of chloride by the amino acid carboxylate in organic solvents, e.g. ethanol, tetrahydrofuran or dioxan). Cleavage with strong acids, e.g. liq. HF, TFMSA
4-Benzyloxybenzyl alcohol (Wang) resin		Active esters or carbodiimides are used for direct attachment. Cleavage proceeds smoothly with 95% TFA. Used routinely in batch Fmoc chemistry.
2,4-Dialkoxybenzyl alcohol resin (SASRIN)		Attachment via activated esters or carbodiimide. Cleavage occurs by 0.5–1% TFA in DCM or acidic hexafluoroisoporpanol in DCM.
4-(Hydroxymethyl)-phenylacetamidomethyl (PAM) resin		First amino acid is introduced by carbodiimide. It is compatible with the Boc/Bzl protection scheme, and is cleaved with strong acids, e.g. TFMSA, HBr/TFA.
Benzhydrylamine (BHA) resin		Direct attachment of first amino acid. Cleavage via acidolysis with HF. TFMSA, HBF$_4$?TFA. Suitable for Boc/Bzl protection scheme.
Hydroxycrotonoyl aminomentyl (Hycram) resin		Completely stable toward acids or bases. Orthogonal to other types of protecting groups. Attachment via activated esters and cleavage via Pd(0) catalyst.
Photolabile linkers with R = OH or NH$_2$		Orthogonal to most chemical protecting groups. Attachment with carbodiimide, and cleavage by photolysis.

Notes: DMC, dichloromethane; TFA, trifluoroacetic acid; TFMSA, trifluoromethane sulfonic acid.

cules include sintered polyethylene, cellulose, chitin, silica, controlled pore glass and glass plates.

Controlled pore glass (CPG) is the most common support matrix for SPS of nucleic acids. CPG is ideal in being rigid, chemically inert and non-swellable. These glass beads contain pores of defined diameters (50–100 nm for most synthetic applications) inside which the polymerization reactions take place. The porosity of CPG increases the surface area of the matrix, which improves the efficiency of the process and reduces reagent consumption. The silylation reaction functionalizes glass and introduces a linker (spacer) upon which the 3′-hydroxyl group of the first deoxyribonucleotide is attached.

8.4 PRACTICE OF SOLID PHASE SYNTHESIS AND ITS APPLICATION

SPS is fast, convenient and easily lends itself to mechanization (automation). The process is applicable to the *in silico* preparation of bio-chips. The utilization of SPS in a continuous-flow mode using solid support-filled columns provides superior and efficient automated technique for the synthesis of biomacromolecules. Major advantages lie in the reduced reagent and solvent consumption, and in the very short coupling cycles (1–2 min for TentaGel polymers of size 8 μm). The reaction progress may be monitored on a real-time basis by recording the conductivity of the solution. Chemically modified silica gel and kieselguhr/polydimethylacrylamide hybrid materials have been successfully applied to automated continuous-flow synthesis (Dryland and Sheppard, 1986). TantaGel-type polymers and CPG are appropriate solid support materials for continuous-flow syntheses (Bayer, 1991) as they are chemically and physically inert and also stable toward moderate pressure. Their uniform spherical shape and homogenous swelling/non-swelling behavior are further advantages. SPS has been developed into automatic peptide/nucleotide syntheses and generalized to organic syntheses. Solid phase synthesis also provides a technique (combinatorial) to rapidly produce large, organized collections of compounds called libraries. This approach to synthetically produce molecular diversity is termed combinatorial chemistry of which SPS of peptides/nucleotides/glycosides contributes prominently to its development and its being. The practice and application of SPS will be considered.

8.4.1 Oligo- and polypeptide synthesis

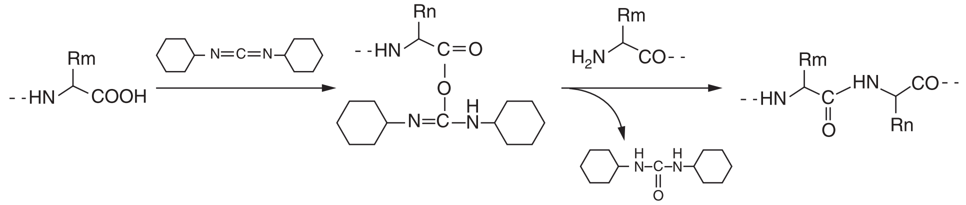
Peptide syntheses are complicated by the fact that nine of the amino acids have functional groups in the side chains that need to be selectively protected. These include amino of Lys, hydroxy of Ser and Thr, phenol of Tyr, thiol of Cys, carboxyl of Asp and Glu, imidazole of His and guanidino of Arg. The thioether group of Met usually does not cause severe complications during peptide synthesis. Side reactions that occur during deprotection reactions with Met-containing peptides are much more problematic, such as partial S-demethylation (base), S-alkylation (acid) or S-oxidation. However, methionine sulfoxide formation has been suggested as a protecting measure for the thioether moiety. The sulfoxide group can be reduced with thioglycolic acid. S-Methylation to give the sulfonium salt has also been recommended as a protecting measure. From a chemical point of view, the ω-carboxyamide groups of Asn and Gln are rather unreactive functionalities that do not require further protection. However, undesired nitrile formation from the carboxamide is prominent under the conditions of carbodiimide couplings. Furthermore, cyclization reaction of Asn with release of ammonia and conversion of Gln into pyroglutamyl peptide of N-terminal Gln has been reported. *N*-Methoxybenzyl or the *N*-diphenylmethyl protecting group can be introduced by reacting with benzyl or diphenylmethyl alcohol respectively. These protecting groups can be cleaved under acidolytic conditions.

Histidine is one of the most problematic building blocks in peptide synthesis. Imidazole ring of His is easily *N*-acylated. Undesired side reactions such as racemization, formation of cyclic imidazolides and *N*-guanylations of the imidazole ring have been observed. These problems may be avoided by reversible masking of the imidazole function such as benzyl (Bzl), 2,4-nitropheyl (Dnp), benzyloxymethyl (Bom), tert-butoxycarbonyl (Boc), triphenylmethyl (Trt), diphenylmethyl (Dpm), pyridyldiphenylmethyl (Pdpm), 4-toluenesulfonyl (Tos), allyl (Al) and allyloxymethyl (Alom).

Orthogonal protection is required for the diamino carboxylic acid (Lys) and aminodicarboxylic acids (Asp and Glu). The ε-amino group of Lys is more basic and more nucleophilic than the α-amino group, and can be converted into the ε-imine with one equivalent of benzaldehyde in the presence of LiOH. The treatment with acid subsequently releases the ε-amino group. Regioselective acid-catalyzed esterification (e.g. with benzyl alcohol) at the ω-carboxyl group can be achieved using H_2SO_4.

The structural requirement of cysteine-containing peptides, where the Cys is needed in its free form while other peptides require regioselective intra- or intermolecular disulfide bridge formation, complicates peptide synthesis. The high nucleophilicity, the ease of oxidation and the acidic character of the cysteine thiol group require selective, semipermanent blocking of this functional group during all synthetic operations. S-Benzyl and derivatives have been introduced as the thiol protecting groups. The deblocking can be achieved by either reduction (sodium in liquid ammonia), and acids (HF, thallium trifluoroacetate). Triphenylmethyl group (Tr) has received much attention as it may be cleaved by silver salt, mercury (II) salts, iodine or trifluoroacetic acid (TFA). Disulfide derivatives (form unsymmetrical disulfide with ethylsulfanyl, tert-butylsulfanyl, 3-nitro-2-pyridylsulfanyl) also serve as protecting groups and can be selectively reduced with 2-mercaptoethanol or phosphines.

In principle, peptides can be synthesized either by stepwise chain assembly, starting preferentially from the C-terminus, or by the condensation of peptide segments (Gross and Meienhofer, 1979; Bodanszky and Bodanszky 1994; Lloyd-Williams *et al.*, 1997). Carbodimides are used for the stepwise assembly of peptides using urethane (Z, Boc, Fmoc)-protected amino acids in peptide synthesis as well as for segment condensations. Initially *N*,*N*′-dicyclohexyl carbodiimide (DCC) was used, but since then several carbodiimide derivatives have been substituted with different tertiary amino and quaternary ammonium groups such as *N*-ethyl-N′-(3-dimethylaminopropyl)carbodiimide (EDC) and 1-cyclohexyl-3-(3-trimethylammoniopropyl)carbodiimide (WSCI) to increase solubility in aqueous solutions. During coupling, the carbodiimide is converted into the corresponding urea derivative, which in the case of *N*,*N*′-dicyclohexyl urea precipitates from the reaction solution. For example, mechanistically the carboxylic anion adds to the protonated carbodiimide with formation of a highly reactive *O*-acylisourea, which reacts with the amino component to produce the peptide and urea derivative:



The application of additives was investigated in order to suppress or diminish side reactions (*N*-acyl urea formation and racemization). *N*-Hydroxysuccinimide (HOSu), 1-hydroxybenzotriazole (HoBt), and ethyl 1-hydroxy-1H-1,2,3-triazole-4-carboxylate are potential additives in the DCC-based coupling synthesis. Solid-phase peptide synthesis (SPPS) is a variant of the linear (stepwise) coupling of amino acids in the C→N direction using two major protection groups; Boc/Bzl (*tert*-butoxycarbonyl/benzyl) and Fmoc/tBu (/9-fluorenylmethoxycarbonyl/*tert*-butyl). The synthetic scheme for peptides on a polymer (Atherton and Sheppard, 1989; Fields, 1997) is illustrated in Figure 8.1.

8.4.1.1 Protection schemes. Each of the repetitive synthesis cycles (deblocking–coupling) is initiated by selective cleavage of the N^{α}-protecting group. The anchoring group (linker) and the semipermanent side chain protecting groups must be

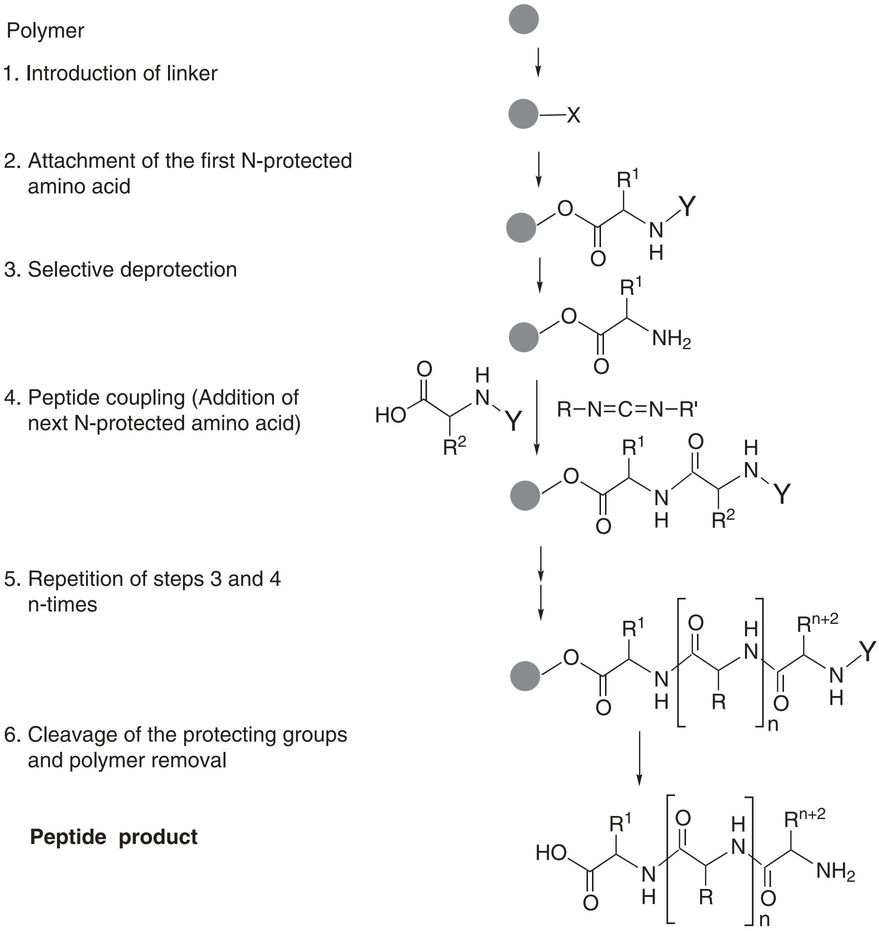

Polymer
1. Introduction of linker
X
2. Attachment of the first N-protected amino acid
R1
O
N
Y
H
3. Selective deprotection
NH2
4. Peptide coupling (Addition of next N-protected amino acid)
HO
R2
R−N=C=N−R'
5. Repetition of steps 3 and 4 n-times
Rn+2
R
n
6. Cleavage of the protecting groups and polymer removal
Peptide product

Figure 8.1 Process of solid-phase peptide synthesis. In the process, X = linker which is later incorporated into the polymer matrix; Y = temporary protecting group; R's = amino acid side chains protected with semipermanent protecting groups if necessary. Carbodiimide is used as the coupling reagent

correctly chosen with respect to the chemical properties of the temporary N^{α}-protecting group, which is usually held constant for all amino acids used throughout the synthesis. In SPPS, either the acid-labile *tert*-butyloxycarbonyl (Boc) or the base-labile fluorenyl-9-methyloxycarbonyl (Fmoc) group, are applied preferably as temporary protecting groups.

- Boc/Bzl protecting groups scheme (Merrifield, 1963):
  The *tert*-butyloxycarbonyl (Boc) group is used as a temporary N^{α}-protector in combination with benzyl-type semipermanent protection, as Boc is usually cleaved with 20–50% TFA. In order to safeguard the stability of the side-chain protection, benzyl-type groups with electron acceptors may be applied (e.g. 2,6-dichlorobenzyl, 2,6-Dcb). Cyclohexyl esters often are applied preferentially for the protection of side-chain carboxyl groups. The semipermanent protecting groups and the benzyhydrylamine (BHA) linker group are cleaved simultaneously by treatment with liquid HF on completion of the synthesis. Scavengers such as anisole, thioanisole, dimethyl sulfide or triisoporpylsilane must be added to the deprotection reaction in order to avoid side reactions of the intermediate carbonium ions. Applicable semipermanent protecting groups are Arg (Tos, Mts), Asp/Glu (Obzl, Ocy), Cys (Acm,

Mob), His (Bom, Dnp, Z, Tos), Lys (2-ClZ), Ser/Thr (Bzl), Trp (For) and Tyr (2-BrZ). Abbreviations used are: Acm, acetamidomethyl; Bom, benzyloxymethyl; Bzl, benzyl; Dnp, dinitrophenyl; For, formyl; Mob, 4-methoxybenzyl; Mts, 2,4,6-trimethylbenzenesulfonyl; Obzl, benzyloxy; Ocy, cyclohexyloxy; Tos, tosyl; Z, benzyloxycarbonyl; 2-BrZ, 2-bromobenzyloxycarbonyl; 2-ClZ, benzyloxycarbonyl.

- Fmoc/tBu protecting groups scheme (Chan and White, 2000):
  These procedures makes use of the base lability of the fluorenyl-9-methyloxycarbonyl (Fmoc) group with two-dimensional orthogonality. Fmoc is cleaved by base-catalyzed elimination where the secondary amine (piperidine) also traps the dibenzofulvene initially formed in the reaction. The semipermanent side-chain protecting groups are mostly of the *tert*-butyl type, and can be cleaved under relatively mild reaction conditions with TFA. Linker moieties displaying comparable acid liability (e.g. Wang resin) are mainly used. The preferred side-chain protecting groups are Asp/Glu (OtBu), Asn/Gln (Trt, Tmb), Cys (Trt), His (Trt), Lys (Boc), Ser/Thr (tBu), Arg (Pmc, Pbf), Trp (Boc) and Tyr (tBu). Abbreviations used are: Boc, *tert*-butoxycarbonyl; tBu, *tert*-butyl; OtBu, *tert*-butoxy; Pbf, 2,2,4,6,7-pentamethyldihydrobenzofuran-5-yl-suflonyl; Pmc, 2,2,5,7,8-pentamethylchroman-6-yl-sulfonyl; Tmb, trimehyoxybenzyl; Trt, tripheylmethyl (trityl)).

8.4.1.2 Chain elongation. In SPPS, complete conversion (coupling) is the basic precondition for the formation of a homogeneous final product, therefore coupling reagents are applied in excess (usually three-fold). This increases the conversion but may also give rise to several undesired side reactions. Peptide couplings in SPPS are usually performed in dichloromethane, sometimes with addition of dimethylformamide, and at ambient temperature with carbodiimide/HOBt (1-hydroxy-1*H*-benzotriazole) as the prevailing method.

8.4.1.3 Peptide cleavage from the resin. As the anchoring group (linker) is usually chosen to be compatible with the peptide synthetic operations, selective cleavage between the C-terminus of the peptide and the solid support occurs upon treatment with reagents that may concomitantly effect partial or complete deprotection of the peptide side chains. In most cases, acidolytic cleavages are used, though other methods such as photolysis for allyl-based linkers (Whitehouse *et al*., 1997) are available. The Boc/Bzl scheme often relies on final cleavage by anhydrous liquid HF at 0°C. This reaction must be carried out in a special apparatus, because HF decomposes glassware. Scavengers such as anisole must be added. Both anchoring groups and all nitrogen- or oxygen-protecting groups based on the *tert*-butyl type or benzyl type are cleaved. Cleavage reactions with TFMSA in TFA can be improved by addition of scavenging sulfur compounds such as thioanisole or methionine. The Fmoc/tBu scheme allows for much milder final deprotection conditions, e.g. with TFA or acetic acid in dichloromethane. Occasionally, the separation of the two steps, linker cleavage and side chain deprotection may offer advantages.

8.4.1.4 Chemical ligation. The size of typical functional domains in protein is ~130 ± 40 (Berman *et al*., 1994). To obtain such proteins by total chemical synthesis for studying structure-function relationships, an approach based on 'chemical ligation'; the stitching together of large unprotected synthetic peptide segments by chemoselective reaction is employed (Dawson and Kent, 2000). Chemical ligation involves the reaction of two unprotected peptide chains having unique functionalities that are mutually reacted with one another, yet unreactive (under the reaction conditions used) with all the other functionalities present in both peptide segments (Figure 8.2). A number of chemistries can be used for the chemical ligation of unprotected peptides such as thioester-forming ligation

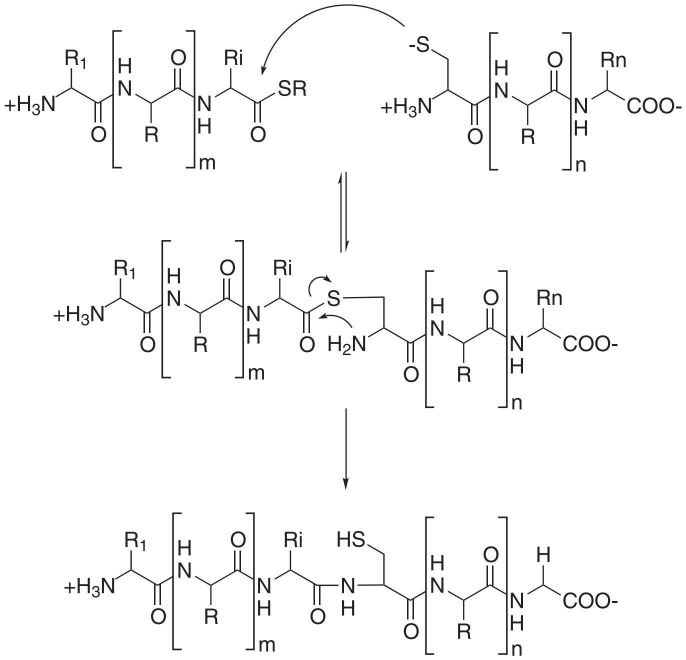


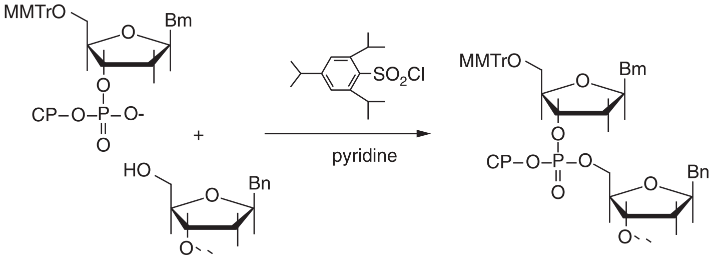
Figure 8.2 Chemical ligation *via* thioester at Cys. The –SH moiety of an N-terminal Cys residue (αCys) undergoes a thiol exchange reaction with the thioester at the C-terminal of the peptide segment. This exchange (transthioesterification) is reversible under the condition used. Uniquely for the αCys, the initial reaction product spontaneously rearranges through a five-membered ring intermediate (S→N acyl shift) to give a native amide bond. This step is irreversible under the conditions used, so the native amide-linked product is formed. An alternative and efficient protein ligation exploiting protein splicing will be described later (Subsection 8.6.1)

(Schnolzer and Kent, 1992), oxime-forming ligation (Rose, 1994) and thiol capture method (Liu and Tam, 1994) and activated amide ligation (Saxon *et al.*, 2000).

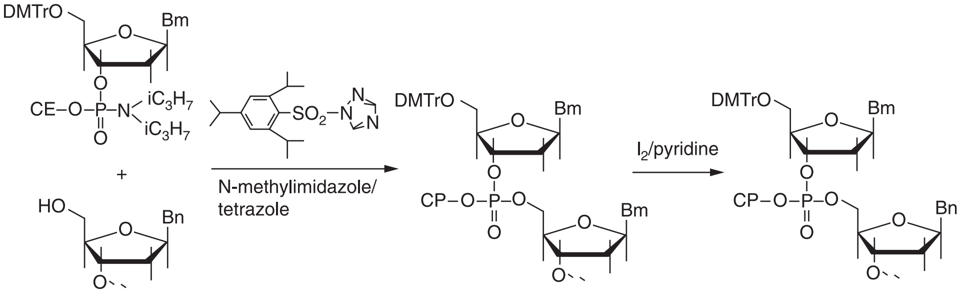
8.4.2 Oligo- and polynucleotide synthesis

In general, the phosphodiester bonds in oligo/polynucleotides are joined between a 3′-hydroxyl group of one nucleoside and a 5′-hydroxyl group of another. The synthesis of genomic DNA can now be readily synthesized by the use of polymerase chain reactions (PCR) (Innis *et al.*, 1999; Mullis *et al.*, 1994). RNA with complementary sequences can be synthesized by the use of DNA-dependent RNA polymerase in the transcription reaction. It has become common practice to refer to all chemically synthesized, single-stranded nucleic acid chains of defined length and sequence as oligonucleotides that can be routinely and reliably synthesized with an automated ‘gene machine’ utilizing SPS process (Balkwell and Rolph, 2002). Chemical synthesis of oligonucleotides is initiated with an introduction of a 3′-phosphate, i.e. 5′-*O*-mono-/dimethoxytrityl (M/DMTr)-(*N*-acylated)-2′-deoxyribonucleotides, which are phosphorylated or phosphitylated at the 3’-hydroxy site. The product after assembly of the oligonucleotide chain is a phosphotriester carrying a protecting group e.g. dichlorophenyl (CP) or 2-cyanoethyl (CE) group (Hardewijn, 2005).

In the phosphotriester synthesis, 5′-*O*-MMTr-(*N*-acylated)-deoxyribonucleotide-3′-*O*-(CP phosphate) is coupled to a 3′-hydroxyl protected deoxyribonucleotide or an existing chain using mesitylenesulfonyl chloride as the coupling agent.


MMTrO
Bm
O
CP–O–P–O-
O
+
HO
Bn
SO2Cl
pyridine
MMTrO
Bm
CP–O–P–O
Bn

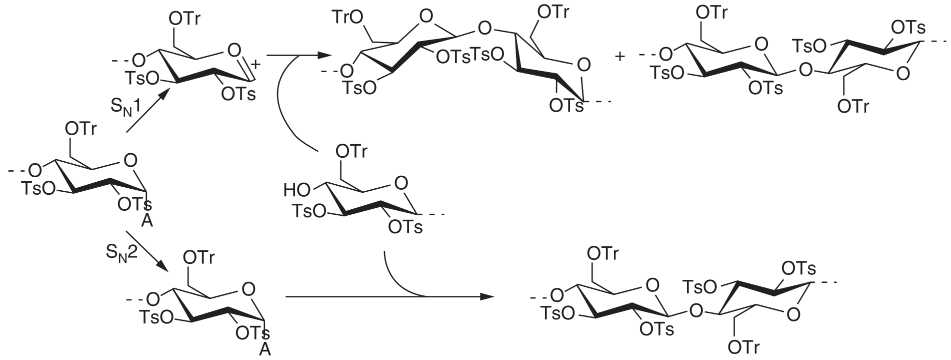
In the phosphite triester synthesis (phosphoramidite synthesis), an efficient coupling reaction takes place between the 3′-hydroxyl group of the chain and 5′-DMTr-(*N*-acylated)-deoxyribonucleoside 3′-*O*-(*N*,*N*-diisopropylamino)phosphite.


DMTrO
Bm
CE–O–P–N
iC3H7
iC3H7
+
HO
Bn
SO2–N
N-methylimidazole/
tetrazole
DMTrO
Bm
CP–O–P–O
Bm
I2/pyridine
DMTrO
Bm
CP–O–P–O
Bn

Tetrazole promotes phosphitylation by protonating the nitrogen atom of phosphaoramidite and avoids a deleterious effect on the protecting DMTr group. The phosphite must be oxidized with iodine to the phosphotriester before proceeding with chain extension. The phosphoramidite synthesis is highly efficient and has been adapted into SPS (Figure 8.3) for DNA synthesizer (the gene machine) and *in silico* synthesis of DNA chips (Schena, 2000).

8.4.3 Oligo- and polysaccharide synthesis

Chemical synthesis of oligo/polysaccharides commonly involves the activation of the anomeric hydroxyl group and displaces the activated group with the acceptor saccharide to form new glycosidic bonds. One issue confronting glycoside synthesis is the selection of orthogonal protecting groups and their selective manipulation during synthesis. Another issue is the control of anomeric configuration of the glycosidic bond formed because the reaction can occur via either an S_N1 or S_N2 mechanism:


OTr
TsO
OTs
A
SN1
SN2
+
TrO
OTs
OTr
HO
TsO
OTs
+
OTr

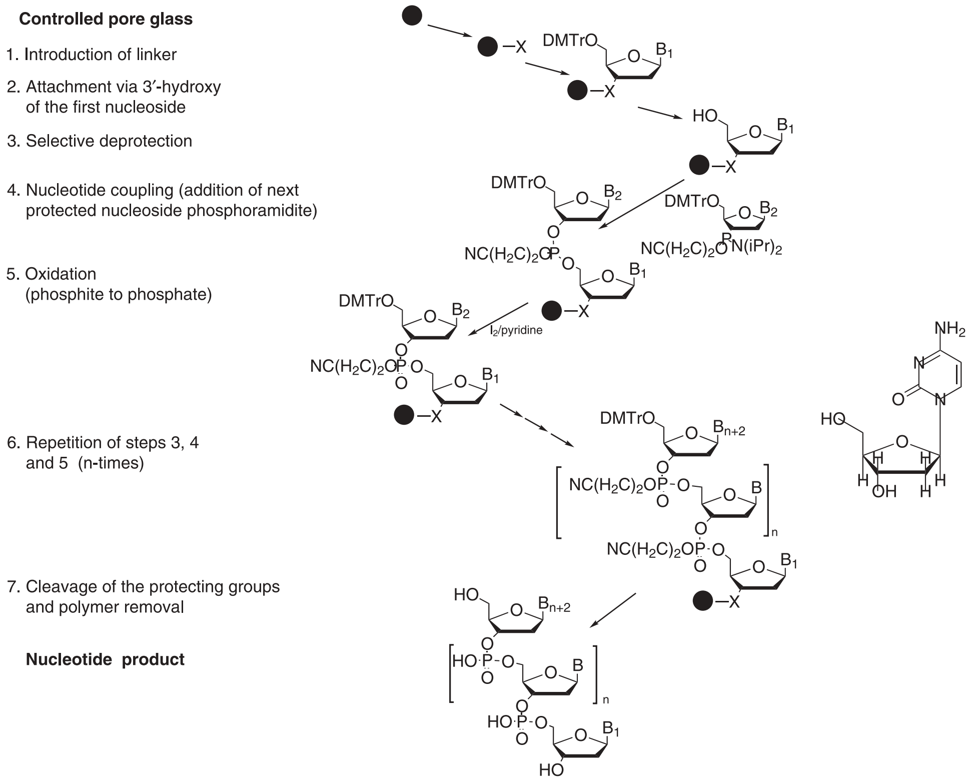

Controlled pore glass
1. Introduction of linker
2. Attachment via 3′-hydroxy of the first nucleoside
3. Selective deprotection
4. Nucleotide coupling (addition of next protected nucleoside phosphoramidite)
5. Oxidation (phosphite to phosphate)
6. Repetition of steps 3, 4 and 5 (n-times)
7. Cleavage of the protecting groups and polymer removal
Nucleotide product
–X
DMTrO
B1
HO
B2
NC(H2C)2O
N(iPr)2
I2/pyridine
Bn+2
B
n
HO
NH2

Figure 8.3 Process of solid-phase phosphoramidite nucleotide synthesis. In the process, X = linker which is later incorporated into the polymer matrix; DMTr = dimethoxy trityl protecting group which is removed by treatment with acid to liberate the hydroxyl group for subsequent phosphitylation reaction. B's = protected purine and pyrimidine bases. Phosphite is oxidized with iodine in the presence of esterification agent to phosphotriester before the chain extension or it is oxidized at the completion of chain extension as shown

The anomeric configuration of the activated saccharide does not ensure the anomeric configuration of the product, which can be greatly influenced by the protecting groups used. Acyl protecting groups at C-2 can strongly direct the *trans* configuration at C-1 (formation of dioxocarbonium ion intermediate in S_N1) but *cis* configuration at C-1 (neighboring group participated double displacement in S_N2). In general, α-1,2-*cis*-glycosides, such as α-D-glucosides and α-D-galactosides, can be formed in the displacement of glycosyl halides, thioglycoside or trichloroacetimidates (without C-2 effect in a nonpolar solvent). β-1,2-*trans*-Glycosides, such as β-D-glucosides and β-D-galactosides, can be obtained by using acyl protecting groups at C- and polar media to favor $S_N1$displacement with the formation of dioxocarbonium intermediate.

Most of these synthetic reactions utilizing the activated anomeric hydroxyl displacements have been applied to SPS of oligosaccharides using polystyrene-based resins (Rademann *et al.*, 1998) and TentaGel (St Hilaire and Meldal, 2000), as illustrated in Figure 8.4.

Enzymatic approaches to glycoside synthesis may circumvent these problems. Enzymes (i.e. nucleoside glycosyl transferases) are stereospecific and regioselective catalyzing glycosyl transfer reaction under mild conditions. Extensive protection-deprotection schemes are unnecessary and the control of anomeric configuration is managed.

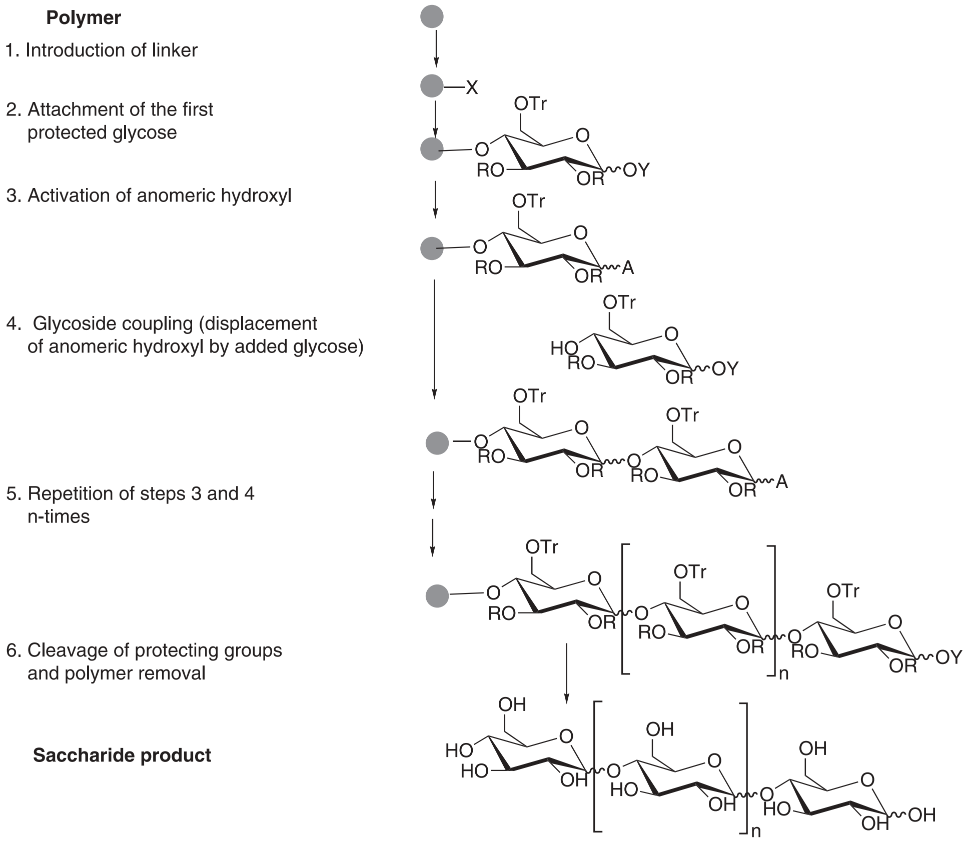
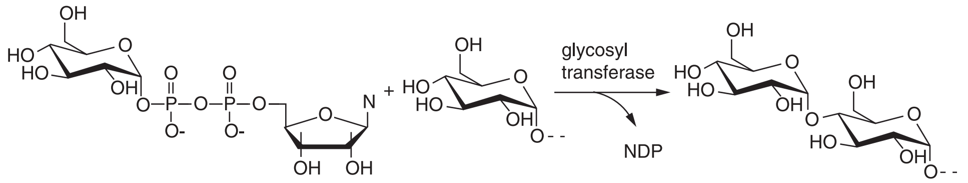

Polymer
1. Introduction of linker
2. Attachment of the first protected glycose
3. Activation of anomeric hydroxyl
4. Glycoside coupling (displacement of anomeric hydroxyl by added glycose)
5. Repetition of steps 3 and 4 n-times
6. Cleavage of protecting groups and polymer removal
Saccharide product
X
OTr
O
RO
OR
OY
A
HO
OH
n

Figure 8.4 Process of solid-phase glycoside synthesis. In the process, X = linker which is later incorporated into the polymer matrix; A = activating group for anomeric hydroxyl; Y = temporary anomeric hydroxyl protecting group; Tr = trityl, and R = semipermanent protecting group for secondary hydroxyls. The wavy bond represents either α- or β-anomeric configuration


glycosyl transferase
NDP
OH
HO
O
P
O-
N
O- -
+

Glycosyl transferases, which utilize dinucleotide glycoses as the glycosyl donor in the biosynthesis of glycans, have been used in the glycoside synthesis (Koeller and Wong, 2000). Their application to the automated SPS of saccharides is conceptually simple (Sears and Wong, 2001); i.e. via immobilized saccharide chain or immobilized enzyme approaches (Figure 8.5).

Glycosidases can be coaxed to synthesize saccharides. Similar attempts have been made with proteases for the peptide synthesis. The equilibrium of a glycosidase-catalyzed reaction normally lies on the side of the thermodynamically more stable cleavage products. The large amount of enzyme required and the low reaction rate are drawbacks of this process. The saccharide competes with the water molecule as the acceptor for the glycosyl moiety. Different manipulations are possible to shift the equilibrium in favor of

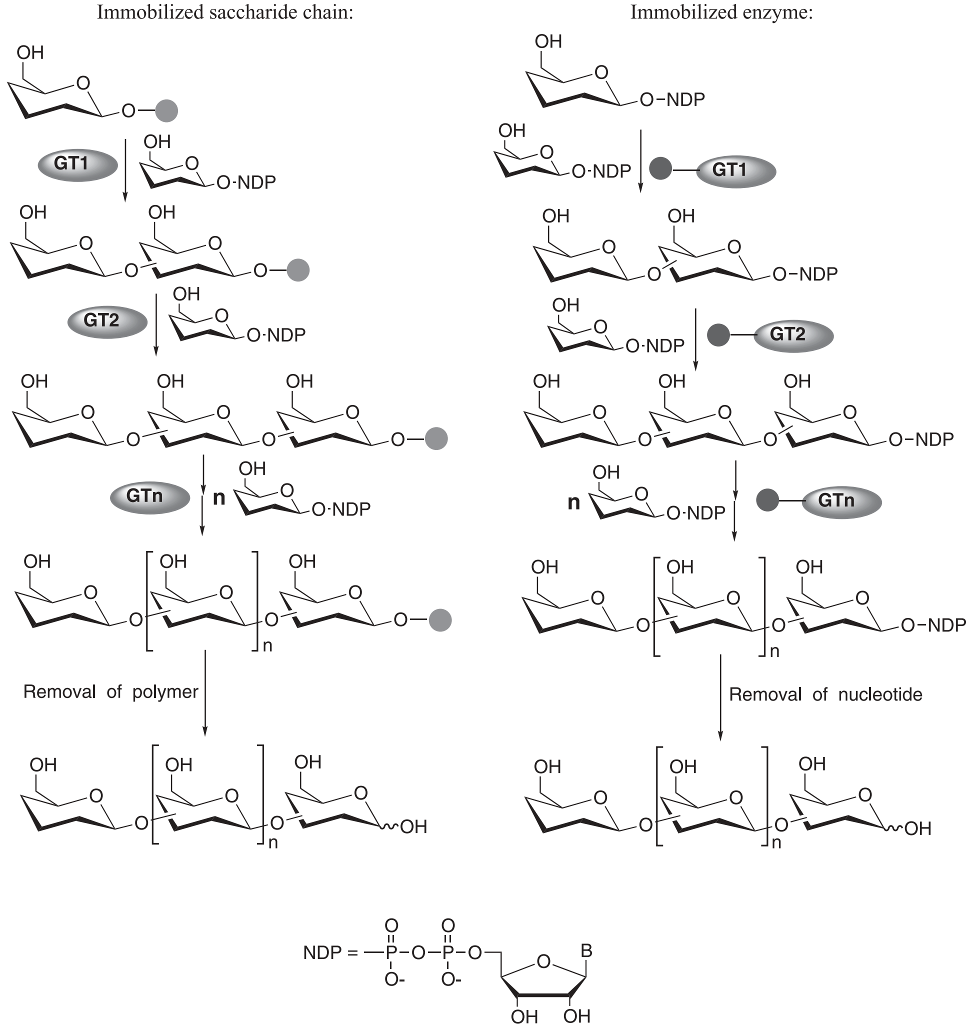


Figure 8.5 Approaches to enzymatic SPS of saccharides. In the immobilized saccharide chain approach, the growing saccharide chain is attached to the polymer support *via* a linker. The free glycosyl transferase (GT) enzymes and nucleotide substrates (NDP-glycoses) are added stepwise to the solid phase reactor. NDP = GDP for Fuc and Man; and UDP for Glc, Gal, Xyl, NAcGlc, NacGal, NacMur, and GlcA (sialic acid uses CMP). In the immobilized enzyme approach, the reaction mixture containing the growing saccharide chain and nucleotide substrates are processed through series of reactors packed with immobilized GT's in sequence. For clarity, secondary hydroxyl groups of pyranose rings are not shown and the positions of the glycosidic bonds are not specified

glycoside bond formation. In the equilibrium controlled approach, the anomeric hydroxyl component contains a free hydroxyl group. Addition of water-miscible organic solvents to the aqueous reaction mixture is a viable option. The kinetically controlled approach is generally limited to glycosidases that rapidly form a glycosyl enzyme intermediate. The glycosyl transfer can be positively manipulated by varying the kind of leaving group. The activated anomeric hydroxyl such as halide or thioglycoside is employed. It is noted that the course of kinetically controlled glycosidase catalyzed synthesis can be more efficiently influenced than the equilibrium controlled approach. Another approach involves the design of glycosynthases (mutant glycosidases), which carry out transglycosylation reactions using readily available glycosyl donors (Mackenzie *et al.*, 1998).

8.5 COMBINATORIAL SYNTHESIS

Combinatorial chemistry concerns with strategies and processes for the rapid synthesis and analysis of large, organized collections of compounds called libraries (Bannwarth and Felder, 2000; Fassina and Miertus, 2004; Fenniri, 2000; Nefzi *et al.*, 1997; Liu and Schultz, 1999). The combinatorial chemistry approach has two phases: i) making a library; and ii) finding the active compound. Thus combinatorial synthesis refers to the process of producing collections of molecularly diverse compounds that can be used for rapid screening for biological activity. The library refers to mixtures of compounds, while collections of single compounds are often called arrays.

Instead of the conventional method of synthesizing individual trinucleotides for codon analysis, it is possible to couple a mixture of all 4 nucleotides to another mixture of 4 nucletotides to produce 16 dinucleotides. If these newly formed 16 dinucleotides are reacted with a mixture of 4 additional nucleotides, 64 trinucleotides that correspond to all possible codons are obtained. The mixture of 64 trinucleotides is called a combinatorial library (nucleotide library). In this combinatorial synthesis, the number of products (P) increases exponentially according to

$$\mathrm{P} = V^n$$

where V is the number of building blocks (variables) and n is the positions to be varied (reaction steps if each step involves one bond formed). To establish the sequences of the trinucleotides in the synthetic mixture with respect to the encoded amino acids (say screening AUG for Met), a procedure called deconvolution is employed. Imagine constructing the library with only 3 nucleotides by leaving out one specific (known) residue. This leads to a library of only 48 trinucleotides after subsequent couplings with 4 nucleotides twice. If this library is inactive (e.g. A is initially omitted and failure to form a complex with methionyl-tRNA), we have learned that position 1 (from 5′-end) of tripeptide has the residue (e.g. A) that was omitted from the synthesis. This process is continued methodically omitting one nucleotide after another to define the best residue in positions 2 and 3 respectively (e.g. failure of complex formation with methionyl-tRNA, if U and G are omitted in the second and third cycles of syntheses respectively). The procedure establishes that the trinucleotide, AUG encodes for Met. This mathematical principle is the parallelism advantage: $4 + 4 + 4 = 12$ (linear) versus $4 \times 4 \times 4 = 64$ (exponential) that underlies the principle of combinatorial chemistry.

Combinatorial synthesis has prompted an interest in rapid synthetic methods, particularly in the area of SPS and its application to the high-speed automated synthesis (Seneci, 2000). In solid-phase combinatorial synthesis, reagents can be used in excess without separation problems to attain complete conversion. Facile purification and automation provide further advantages. Two approaches for combinatorial synthesis of oligonuclotides/peptides/saccharides will be considered.

8.5.1 Parallel synthesis

Multiple nucleotide/peptide/saccharide synthesis refers to the simultaneous (parallel) synthesis of a multitude of nucleotide/peptide/saccharide sequences, irrespective of the chain length and monomer composition. There are differences between the variants of multiple syntheses with respect to the polymer support used. For example, in the multipin peptide synthesis, peptide assembly is performed using amino functionalized polyethylene rods (pins) and coupling reactions on an array of pins are performed in parallel in an array of microtiter plate wells. An analogous spot synthesis uses a planar sheet of cellulose or cotton-wool as the polymeric support. The first monomer component is joined via an ester bond and a linker molecule to the hydroxyl groups of cellulose. Residual (unoccupied) hydroxy functions of cellulose must be deactivated prior to the coupling reactions. Chemically functionalized polystrene-polyethylene films and glass plate may also be used as solid support materials in the spot combinatorial synthesis.

The light-directed, spatially addressable parallel combinatorial synthesis is based on the combination of photolithographic techniques with solid-phase synthesis using photolabile protecting groups. For example, photolithographic peptide synthesis can perform on functionalized glass plates (Fodor *et al*., 1991) using photolabile amino protecting group, 6-nitroveratryloxycarbonyl (Nvoc). Using a lithographic mask, photodeprotection (with light of wavelength 365 nm) of the terminal amino group exposes the deprotected sites for the coupling of the next Nvoc-protected amino acid. The process is repeated with suitable masks to control the irradiated deprotection sites, subsequent coupling reactions and therefore the sequences of the peptide chains (Figure 8.6). This positionally addressable spatial array prepares each compound immobilized on a separate spot on a surface and allows the *xy* coordinates of the spot to be related to the structure.

8.5.2 Mixture synthesis

Combinatorial libraries of nucleotides/peptides or diversomers (organic compounds) are obtained by mixture syntheses, which are designed to produce a mixture of nucleotides/peptides or diversomers with an optimum degree of heterogeneity in an efficient manner. The application of suitable testing methods allows the identification and isolation of the desired compound with a desired biological activity. Mixture synthesis in its basic form uses a mixture of reagents with a predefined ratio of the building blocks (diversity elements). One of the commonly adapted techniques. known as split-and-combine (split-and-pool) combinatorial synthesis, provides a library in the form of a spatially resolved compound mixture (one bead-one compound) as illustrated for the nucleotide synthesis in Figure 8.7.

The advantages gained in the combinatorial synthesis and in the testing of the compound mixture are compensated by the necessary deconvolution in order to identify an active compound. The split-and-combine synthesis produces a library wherein each bead has a single compound attached, encoding (i.e. derivatization of each bead with a ‘tag’ that contains information on the structure attached to the bead) of the library provides one means of identifying the code (synthetic history) rather than the compound directly. The split-and-combine synthesis is carried out as usual. After each ‘react’ step (before ‘combine’) a coding reaction is performed on the bead at a trace level (~1%). Usually the code (tag) is linked with photolabile linker so that the tag can be removed without deleterious effect on the library members (e.g. nucleotides/peptides/saccharides) synthesized. The tag, such as chloroaryl derivatives with varying CH_2 chain can be silylated and identified readily by GC. The identification of the code defines the synthetic steps applied to the bead and therefore the compound (library member) attached to it (Still, 1996).

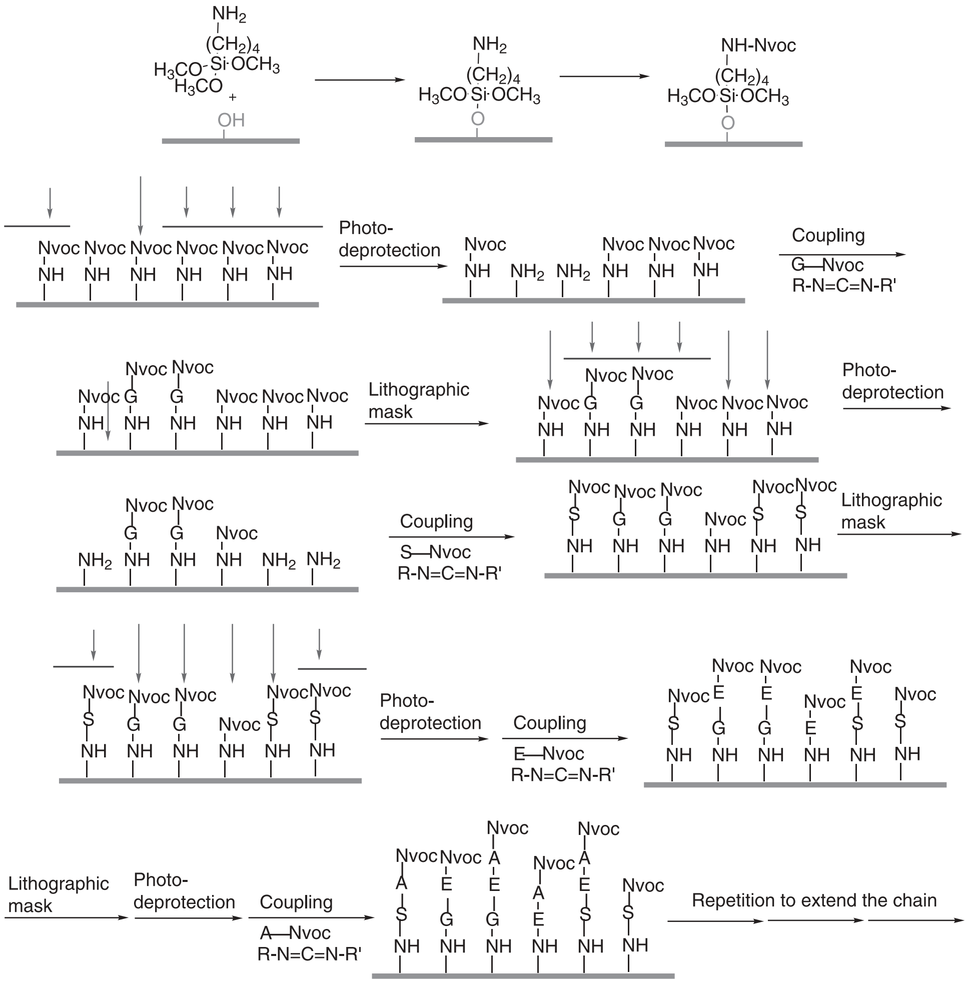
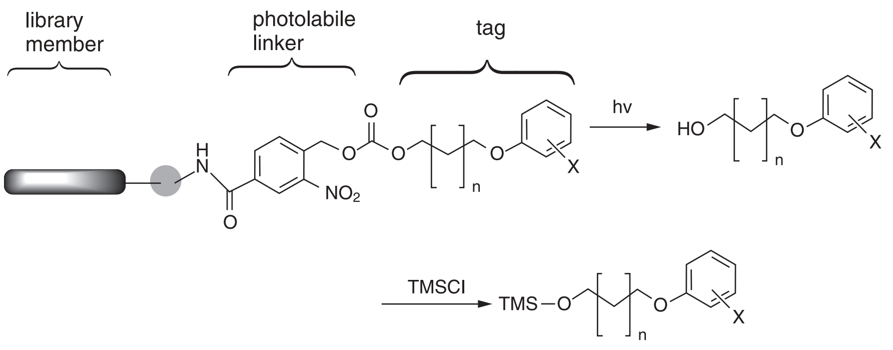

NH2
(CH2)4
H3CO-Si-OCH3
H3CO
OH
NH-Nvoc
Nvoc
NH
Photo-
deprotection
Coupling
G—Nvoc
R-N=C=N-R'
Lithographic
mask
S—Nvoc
E—Nvoc
A—Nvoc
Repetition to extend the chain

Figure 8.6 Schematic representation of photolithographic spatially addressable multiple peptide synthesis. The glass plate is amino functionalized with 4-aminobutyltrimethoxysilane. The carbodiimide promoted peptide synthesis is carried out using Nvoc as the photolabile N^{α}-protecting group. The suitable lithographic masks control the sites of chain extensions in each synthetic step by specifically photo-deprotecting exposed Nvoc-αNH_2 groups (X = Cl, F)


library
member
photolabile
linker
tag
hv
TMSCl
HO
TMS—O
NO2
X
n

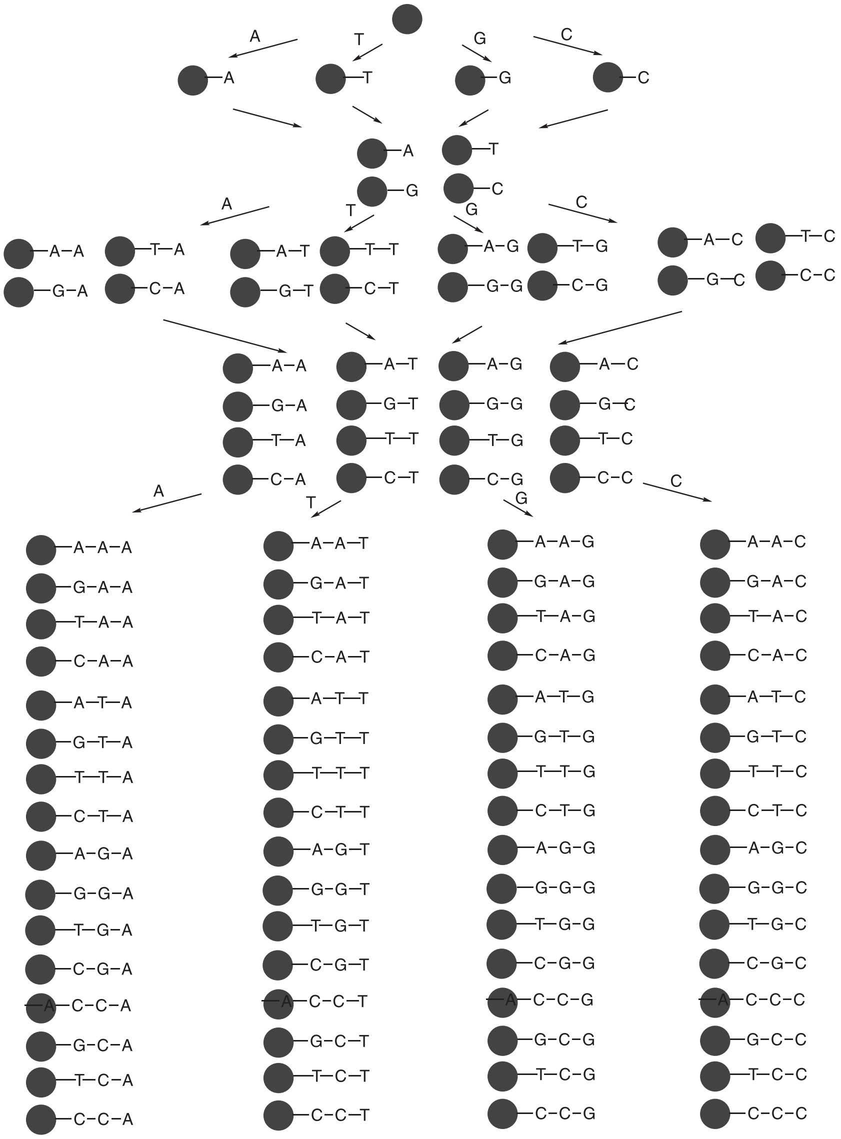


Figure 8.7 Schematic representation of split-and-pool combinatorial nucleotide synthesis. An example of split-and-pool (split-and-combine) synthesis of trideoxyribonucleotides complementary to codons by the solid phase phosphoramidite nucleotide synthesis is illustrated. Each cycle in the synthesis includes split (equal molar split), react (deprotection, coupling with 4 protected nucleotides and oxidation) and combine steps. After 3 cycles of 4 reactions (corresponding to coupling reactions with 4 nucleotides per cycle), the final combined library consists of a mixture all 64 trinucleotides

Suitable testing methods in combination with synthetic nucleotide/peptide/saccharide libraries enable investigations to be made on diverse biochemical issues in a more efficient manner. OSDB (http://www.orgsyn.org/) that compiles organic reactions and experimental procedures is a useful organic synthetic resource site.

8.6 BIOCHEMICAL POLYPEPTIDE CHAIN LIGATION

Proteins have been the major focus of biochemical research in an attempt to fully understand their innate physiological function and to harness that function for diversified applications. Site-directed mutagenesis is invaluable to the study of structure-functional relationships of proteins (Knowles, 1987). However, this approach is limited by the functional groups present in the 20 genetically encoded amino acids. Many studies require access to protein molecules that are either impossible or insufficient to obtain from biological systems, including natural posttranslationally modified proteins and proteins possessing unnatural amino acids. The former is essential for investigating the *in vivo* function of proteins (functional proteomics) and the latter is useful to application and structure-functional studies of proteins. Various chemical and biochemical approaches have been developed to synthesize specifically modified proteins (Hahn and Muir, 2005), some of which will be described.

Chemical ligation (Dawson and Kent, 2000) can be employed to incorporate modified/unnatural amino acids (referred to as unnatural amino acids) by ligation of polypeptide chain containing the desired amino acid derivatives. The chemical ligation is compatible with naturally occurring as well as some unnatural amino acids and ideally suited to protein semisynthesis. The only requirement is that the polypeptide fragments contain one or the other of the necessary reactive groups, either αCys (N-terminal) or thioester (C-terminal). The synthetic polypeptide chain containing the unnatural amino acid is joined with the other synthetic or recombinant polypeptide chain via transthioesterification followed by S→N acyl transfer (Figure 8.2). Biochemical protein chain ligation can also be accomplished by exploiting the process of protein splicing (subsection 13.5.6), known as expressed protein ligation (Muir, 2003).

Protein splicing (Paulus, 2000) is a posttranslational process in which a precursor protein undergoes a series of intramolecular rearrangements and internal reactions that result in the precise removal of an internal segment (intein) and ligation of the two flanking portions (exteins). The exploitation of protein splicing gives rise to two approaches for the semisynthesis that allows synthetic and/or recombinant polypeptides to be chemoselectively and regioselectively joined together (Muir, 2003):

8.6.2.1 Expressed protein ligation by utilizing intact inteins. Expressed protein ligation (EPL) uses a tagged intein with the mutated N-terminal Asn residue (e.g. Asn → Asp), which upon treatment with a thiol reagent generates the N-terminal segment of peptide chain as N-extein. The subsequent reaction of the N-extein with αCys-containing peptide (C-extein) joins the two peptide chains (Severinov and Muir, 1998), as illustrated in Figure 8.8.

8.6.2.2 Protein trans-splicing by employing split inteins. The procedure, also known as intein-mediated protein legation (IPL), is based on the observation that inteins can be cut into two individually inactive pieces, which regain the splicing activity when noncovalently combined/associated. The two halves of inteins are fused to respective exteins. The reconstituted split intein then mediates a normal protein splicing reaction (Yamazaki *et al*., 1998), as shown in Figure 8.9. The two polypeptide chains can be either synthetic and/or recombinant. The unnatural amino acid(s) is introduced via SPPS into the synthetic polypeptide chain that is, in turn, ligated to form a full length protein.

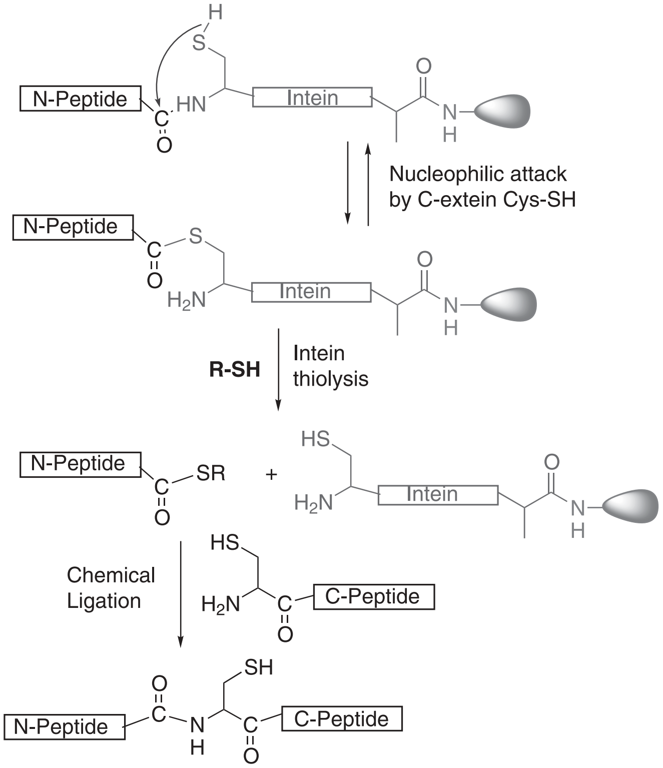

N-Peptide
Intein
Nucleophilic attack
by C-extein Cys-SH
R-SH
Intein
thiolysis
SR
Chemical
Ligation
C-Peptide

Figure 8.8 Expressed protein ligation. The final step of protein splicing by the intein is inactivated by the mutation of the C-terminal Asn to Ala. Proteins expressed as in-frame N-terminal fusions to such mutant inteins can be cleaved by thiols to give corresponding protein (N-peptide) thioester derivatives. The tagged inteins are removed. The N-peptide thioesters can then react with an αCys-containing peptide (C-peptide) *via* chemical ligation to join the two peptides

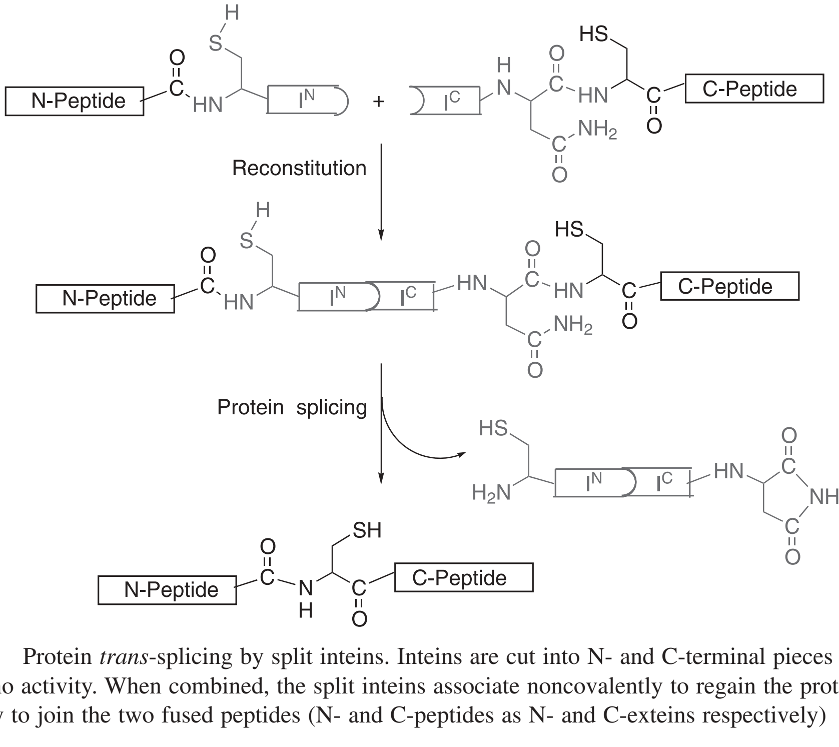

N-Peptide
I^N
I^C
C-Peptide
Reconstitution
Protein splicing

Figure 8.9 Protein *trans*-splicing by split inteins. Inteins are cut into N- and C-terminal pieces (I^N and I^C) that have no activity. When combined, the split inteins associate noncovalently to regain the protein splicing activity to join the two fused peptides (N- and C-peptides as N- and C-exteins respectively)

8.7 REFERENCES

Atherton, E. and Sheppard, R.C. (1989) *Solid Phase Peptide Synthesis: A Practical Approach*, IRL Press, Oxford, UK.

Balkwell, F. and Rolph, M. (2002) *Gene Machines*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Bannwarth, W. and Felder, E. (eds) (2000) *Combinatorial Chemistry: A Practical* Approach, Wiley-VCH, New York.

Bayer, E. (1991) *Angew. Chem. Int. Ed. Engl.*, **30**, 113.

Berman, A.L., Kolker, E. and Trifonov, E.N. (1994) *Proceedings of the National Academy of Science, USA*, **91**, 4044–7.

Bodanszky, M. and Bodanszky, A. (1994) *The Practice of Peptide Synthesis*, Springer-Verlag, Berlin.

Chan, W.C. and White, P.D. (2000) *Fmoc Solid Phase Peptide Synthesis: A Practical Approach*, Oxford University Press, Oxford, UK.

Dawson, P.E. and Kent, S.B.H. (2000) *Annual Reviews in Biochemistry*, **69**, 923–60.

Dryland, A. and Sheppard, R.C. (1986) *Journal of the Chemistry Society Chemistry Communications*, **1986**, 125.

Ernst, B., Hart, G.W. and Sanay, P. (eds) (2000) *Carbohydrates in Chemistry and Biology*, Wiley-VCH, Weinheim, Germany.

Fassina, G. and Miertus, S. (eds) (2004) *Combinatorial Chemistry and Technology*. 2nd edn, Taylor & Francis, Boca Raton, FL.

Fenniri, H. (ed.) (2000) *Combinatorial Chemistry: A Practical Approach*, Oxford University Press, Oxford, UK.

Fields, G.B. (1997) *Solid-Phase Peptide Synthesis*, Academic Press, New York.

Fodor, S.P.A., Reed, J.L. Pirrung, M.C. *et al.* (1991) *Science*, **251**, 767–73.

Greene, J.J. and Rao, V.B. (eds) (1998) *Recombinant DNA Principles and Methodologies*, Marcel Dekker, New York.

Gross, E. and Meienhofer, J. (1979) *The Peptides: Analysis, Synthesis, Biology*, Academic Press, New York.

Hahn, M.E. and Muir, T.W. (2005) *Trends in Biochemical Sciences*, **30**, 26–34.

Hardewijn, P. (2005) *Oligonucleotide Synthesis: Methods and Applications*, Humana Press, Totowa, NJ.

Howl, J. (2005) *Peptide Synthesis and Applications*, Humana Press, Totowa, NJ.

Innis, M., Gelfand, D. and Sninsky, J. (eds) (1999) *PCR Methods Manual*, Academic Press, San Diego.

James, J.W. (1999) *Tetrahedron*, **55**, 4855–946.

Kates, S.A. and Albericio, F. (eds) (2000) *Solid-Phase Synthesis: A Practical Guide*, Marcel Dekker, New York.

Knowles, J.R. (1987) *Science*, **236**, 1252–8.

Koeller, K.M. and Wong, C.-H. (2000) *Chemistry Review*, **100**: 4465–94.

Kullmann, W. (1987) *Enzymatic Peptide Synthesis*, CRC Press, Boca Raton.

Liu, C. F. and Tam, J.P. (1994) *Journal of the American Chemistry Society*, **116**, 4149–53.

Liu, R.L. and Schultz, P.G. (1999) *Angrew. Chem. Int. Ed. Engl.*, **38**, 36–54.

Lloyd-Williams, P., Albericio, F. and Giralt, E. (1997) *Chemical Approaches to the Synthesis of Peptides and Proteins*, CRC Press, Boca Raton.

Mackenzie, L.F., Wang, Q., Warren, R.A.J. and Withers, S.G. (1998) *Journal of the American Chemistry Society*, **120**, 5583–4.

Merrifield, R.B. (1963) *Journal of the American Chemistry Society*, **85**, 2149–53

Merrifield, R.B. (1985) *Angew. Chem. Int. Ed. Engl.*, **24**, 799.

Muir, T.W. (2003) *Annual Reviews in Biochemistry*, **72**, 249–89.

Mullis, K.B., Ferré, F. and Gibbs, R.A. (1994) *The Polymerase Chain Reaction*, Birkhaüser, Boston, MA.

Nefzi, A., Ostresh, J.M. and Houghten, R.A. (1997) *Chemistry Review*, **97**, 449–72.

Paulus, H. (2000) *Annual Reviews in Biochemistry*, **69**, 447–96.

Pennington, M.W. and Dunn, B.M. (1994) *Peptide Synthesis Protocols*, Humana Press, Totowa, NJ.

Rademann, J., Geyer, A. and Schmidt, R.R. (1998) *Angrew, Chem. Int. Ed. Engl.*, **37**, 1241–5.

Rose, K. (1994) *Journal of the American Chemistry Society*, **116**, 30–4.

Saxon, E., Armstrong, J.I. and Bertozzi, C.R. (2000) *Org. Lett.*, **2**, 2141–3.

Schena, M. (ed.) (2000) *DNA Microarrays: A practical approach*, 2nd edn, Oxford University Press, Oxford, UK.

Schnolzer, M. and Kent, S.B.H. (1992) *Science*, **256**, 221–5.

Sears, P. and Wong, C.-H. (2001) *Science*, **291**, 2344–4.

Seneci, P. (2000) *Solid Phase Synthesis and Combinatorial Technologies*, John Wiley & Sons, New York.

Severinov, K. and Muir, T.W. (1998) *Journal of Biological Chemistry*, **273**, 16205–9.

St Hilaire, P.M. and Meldal, M. (2000) *Angew. Chem. Int. Ed. Engl.*, **39**, 1162–79.

Still, W.C. (1996) *Acc. Chemistry Research*, **29**, 155–63.

Whitehouse, D.L., Savinov, S.N. and Austin, D.J. (1997) *Tetrahedron Letters*, **38**, 7851.

Yamazaki, T., Otomo, T. and Oda, N. (1998) *Journal of the American Chemistry Society*, **120**, 55912.

World Wide Webs cited

Chemical DB for SPS: http://www.accelrys.com/products/chem_databases/databases/solid_phase_synthesis.html.

Combinatorial Reference:	http://vesta.pd.com
OSDB:	http://www.orgsyn.org/

CHAPTER 9

STUDIES OF BIOMACROMOLECULAR STRUCTURES: COMPUTATION AND MODELING

9.1 POTENTIAL ENERGY AND MOLECULAR THERMODYNAMICS

The molecular basis for the formation of the biologically functional structures of biomacromolecules is one of the most fascinating and challenging problems in biochemistry. The basic principle that the information needed to fold biomacromolecule into its native 3D structure is contained within its sequence was shown by an experiment demonstrating the refolding of the denatured (unfolded) ribonuclease into its native structure (Anfinsen *et al.*, 1961; Anfinsen, 1973). Theoretically all the physical and chemical properties of a biomacromolecule, including its 3D structure, can be predicted from an accurate description of the total thermodynamic state of the system at the atomic level. One of the ultimate goals in computational biochemistry (Becker *et al.*, 2001; Tsai, 2002) is to accurately predict/model the 3D structure of a biomacromolecule starting with its sequence (French and Brady, 1990; Leontis and SantaLucia, 1998; Webster, 2000). This goal has not been achieved, though progress has been made toward its realization.

The problem in trying to model the 3D structure of a biomacromolecular is formidable because of the large size and complexity of the system and therefore some simplifications and assumptions are required:

1. *Molecular mechanics*: The accurate predictions of the structure and physical properties for a molecule can be made from an exact quantum mechanical treatment of every atom within the molecular system. However, a simpler molecular mechanical treatment is applied to solve a complex macromolecular system.
2. *Energy minimum*: The native conformation of a macromolecule is the one with the lowest overall potential energy. Therefore energy minimization is applied to search for the native/most stable conformation of a biomacromolecule.
3. In vacuo: All the biomacromolecules function in solvent environments, but the complexity of the system is greatly reduced *in vacuo* by removing all the solvent and ions. Various approximations (e.g. bulk dielectrics, periodic box) have been incorporated into the modeling methods to simulate solvent effects without explicitly including solvent molecules.
4. *Periodic box*: The properties of a population of molecules is represented by the time-average behavior of a single molecule and its associated solvent isolated in a

---

Biomacromolecules, by C. Stan Tsai
Copyright © 2007 John Wiley & Sons, Inc.

periodic (repeating) box. The molecule in the box will eventually sample all the possible conformations accessible to the system. The content of this box is identical to that of all the other boxes, and anything that leaves the box from one direction must simultaneously enter it from the opposite direction.

Computational approaches to potential energy of a molecule may be divided into two broad categories: quantum mechanics (Hehre *et al.*, 1986) and molecular mechanics (Berkert and Allinger, 1982). The basis for this division depends on the incorporation of the Schrödinger equation or its matrix equivalent. It is now widely recognized that both methods reinforce one another in an attempt to understand chemical and biochemical behavior of biomolecules at the molecular level. From a purely practical standpoint, the complexity of the problem, time constraints, computer size and other limiting factors typically determine which method is feasible.

Quantum mechanics (QM) can be further divided into *ab initio* and semi-empirical methods. The *ab initio* approach uses the Schrödinger equation as the starting point with post-perturbation calculation to solve electron correlation. Various approximations are made that the wave function can be described by some functional form. The functions used most often are a linear combination of Slater type orbitals (STO), exp(−ax) or Gaussian type orbitals (GTO), $\exp(-ax^2)$. In general, *ab initio* calculations are iterative procedures based on self-consistent field (SCF) methods. Self-consistency is achieved by a procedure in which a set of orbitals is assumed and the electron–electron repulsion is calculated. This energy is then used to calculate a new set of orbitals and these in turn are used to calculate new repulsion energy. The process is continued until convergence occurs and self-consistency is achieved. However, the term semi-empirical is usually reserved for those calculations where families of difficult-to-solve integrals are replaced by equations and parameters that are fitted to experimental data. Semi-empirical methods describe molecules in terms of explicit interactions between electrons and nuclei and are based on the principles:

- Nuclei and electrons are distinguished from each other.
- Electron–electron and electron–nuclear interaction are explicit.
- Interactions are governed by nuclear and electron charges, i.e. potential energy and electron motions.
- Interactions determine the space distribution of nuclei and electrons and their energies.

For the best result, the molecule being computed should be similar to molecules in the database used to parameterize the method. However, if the molecule being computed is significantly different from anything in the parameterization set, erratic results may be obtained. Semi-empirical calculations have been successful in dealing with organic compounds.

For biomacromolecules to which neither QM nor semi-empirical calculations can be applied effectively, the methods referred to as molecular mechanics can be used to model their structures and behaviors. Molecular mechanics (MM) uses simple algebraic expressions for the total energy of a compound without computing a wave function or total electron density (Boyd and Lipkowitz, 1982; Brooks *et al.*, 1988). MM, where molecular motions are determined by the masses of and the forces acting on atom, is based on the following principles:

- Nuclei and electrons are lumped together and treated as unified atom-like particles.
- Atom-like particles are treated as spherical balls.

- Bonds between atom-like particles are viewed as springs.
- Interactions between these particles are treated using potential functions derived from classical mechanics.
- Individual potential functions are used to describe different types of interactions.
- Potential energy functions rely on empirically derived parameters that describe the interactions between sets of atoms.
- The potential functions and the parameters used for evaluating interactions are termed a force field (FF).
- The sum of interactions determines the conformation of atom-like particles.

Therefore MM energies have no meaning as absolute quantities. They are used for comparing relative strain energy between two or more conformations.

According to classical mechanics, the total energy, **E** of a molecular system includes the kinetic energy, K and the potential energy, V:

$$\mathbf{E} = \mathrm{K} + \mathrm{V}$$

The kinetic energy includes all of the motions of the atoms in the system and the potential energy of a macromolecule is represented by a multidimensional surface consisting of hills and valleys with any particular conformation corresponding to one point on this surface. Since V is temperature-independent, T only affects the kinetic energy term. Thus the energy of the system, **E** derives primarily from the potential energy term at low temperature but mainly from the kinetic energy term at high temperature. In MM where the nuclei and electrons are treated together, the nuclei contribute the mass while the electrons provide the force of interactions between atoms. Since the force, F exerted in the direction, r is related to the mass, m and acceleration, a along a molecular trajectory according to:

$$\mathrm{F} = \mathrm{ma} = -\partial \mathrm{V}/\partial \mathrm{r}$$

Thus the force applied on an atom depends on how the potential energy changes as the distance between interacting atoms changes.

A system at equilibrium where $\mathrm{F} = 0$, and therefore $-\partial \mathrm{V}/\partial \mathrm{r} = 0$, means that the molecule sits at a potential energy minimum. This is the basis of energy minimization that attempts to find the lowest energy conformation of a macromolecule, providing a static state of the system at equilibrium.

The kinetic energy of an atom is related to its velocity, v or its momentum, p by:

$$\mathrm{K} = \tfrac{1}{2}(\mathrm{m}v^2) = \tfrac{1}{2}(p^2/\mathrm{m})$$

This describes the dynamic change in the atomic positions at any time, t. Therefore molecular dynamics is used to simulate the time dependent changes in a molecular system. The potential energy, V of an isolated single molecule depends on the intramolecular interactions. The total intramolecular potential energy, $\mathrm{E}_{\mathrm{total}}$ (V and E are used interchangeably) is the sum of the bonding interactions, E_{b} and the nonbonding interactions, E_{nb} expressed as:

$$\mathrm{E}_{\mathrm{total}} = \sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{E}_{\mathrm{b}} + \mathrm{E}_{\mathrm{nb}})_{\mathrm{i}} \qquad \text{for N number of atoms in the molecule.}$$

The bonding interactions are the covalent bonds that hold the atoms together, and the nonbonding interactions include electrostatic, dipolar and steric interactions. In MM, the potential functions that define the force field (FF) are derived empirically (also known as empirical force field, EFF) based on the actually observed macroscopic properties of macromolecular constituents.

9.2 MOLECULAR MODELING: MOLECULAR MECHANICAL APPROACH

9.2.1 Introduction

The fundamental assumption of MM or its tool, empirical force field (EFF or simply force field, FF) is that data determined experimentally for small molecules/macromolecular constituents can be extrapolated to macromolecules. It is aimed at quickly providing energetically favorable conformations for large systems invariably using computers. An application of computers to generate, manipulate, calculate and predict realistic structures and associated properties of molecular systems is known as molecular modeling (Leach, 1996; Schlecht, 1998; Höltje *et al.*, 2003). It is based on theoretical chemistry methods and experimental data that can be used either to analyze molecules and molecular systems or to predict chemical and biochemical properties (Gundertofte and Jørgensen, 2000; Warshel, 1991). It serves as a bridge between theory and experiment to:

- extract results for a particular model;
- compare experimental results of the system;
- compare theoretical predictions for the model;
- help understanding and interpreting experimental observations;
- correlate between microscopic details at atomic and molecular level and macroscopic properties;
- provide information not available from real experiments.

Thus molecular modeling is primarily a mean of communication between scientist and computer, the imperative interface between human-comprehensive symbolism and the mathematical description of the molecule. The endeavor is made to perceive and recognize a molecular structure from its symbolic representations with a computer. Thus functions of molecular modeling include:

- *Structure retrieval or generation*: Crystal structures of organic compounds can be found in the Cambridge Crystallographic Data Centre (CCDC) database (http://www.ccdc.cam.ac.uk/). Those that do not exist may be generated by 3D rendering software. The 3D structural coordinates of biomacromolecules can be retrieved from the Protein Data Bank (http://www.rcsb.org/pdb/).
- *Structural visualization*: Computer graphics is the most effective means for visualization and interactive manipulation of molecules and molecular systems. Numerous software programs (e.g. Cn3D, RasMol and KineMage, Chapter 5) are available for visualization, management and manipulation of molecular structures.
- *Energy calculation and minimization*: One of the fundamental properties of molecules is their energy content and energy level. Three major theoretical computational methods in their calculation include empirical (molecular mechanics), semiempirical and *ab initio* (quantum mechanics) approaches. Energy minimization results in geometry optimization of the molecular structure.

- *Dynamics simulation and conformation search*: Solving the motion of nuclei in the average field of the electrons is called. Solution to Newton's equation of motion for the nuclei is known as molecular dynamics. Integration of Newton's equation of motion for all atoms in the system generates molecular trajectories. Conformation search is carried out by repeating the process by rotating reference bonds (dihedral angles) of the molecule under investigation to find the lowest energy conformations of molecular systems.
- *Calculation of molecular properties*: Methods of estimating or computing properties (i.e. interpolating properties, extrapolating properties and computing properties). Some computable properties are boiling point, molar volume, solubility, heat capacity, density, thermodynamic quantities, molar refractivity, magnetic susceptibility, dipole moment, partial atomic charge, ionization potential, electrostatic potential, van der Waals surface area and solvent accessible surface area.
- *Structure superposition and alignment*: Computing activities and properties of molecules often involve comparisons to be made across a homologous series. Such techniques require superposition or alignment of structures.
- *Molecular interactions, docking*: The intermolecular interaction in a ligand-receptor complex is important for the understanding of the mechanisms of molecular actions. The molecular interactions and docking are difficult modeling exercises. Usually the receptor (e.g. protein) is kept rigid or partially rigid while the conformation of ligand molecule is allowed to change.

The advent of high-speed computers, availability of sophisticated algorithms and state-of-the-art computer graphics have made plausible the use of computationally intensive methods such as QM, MM and molecular dynamics (MolD) simulations to determine those physical and structural properties most commonly involved in molecular processes. The power of molecular modeling rests solidly on a variety of well-established scientific disciplines including computer science, theoretical chemistry, biochemistry and biophysics. Molecular modeling has become an indispensable complementary tool for most experimental scientific research.

Computer-assisted molecular modeling has rapidly become a vital component of biochemical research. Mechanisms of ligand-receptor and enzyme-substrate interactions, protein folding, protein-protein and protein-nucleic acid recognition and *de novo* protein engineering are a few examples of problems that may be addressed and facilitated by this technology.

Molecular mechanics is a widely used method for generating molecular models for a multitude of purposes in chemistry and biochemistry. The first major reason for the popularity of MM is its speed, which makes it computationally feasible for routine usage. The alternative methods for generating molecular geometry, such as *ab initio* or semi-empirical molecular orbital calculations, consume much larger amounts of computer time, making them more expensive to use. The economy of MM makes studies of relatively large molecules, such as biomacromolecules, feasible on a routine basis. Thus MM has become a primary tool of computational biochemists. Molecular mechanics is relatively simple to understand. The total strain energy is broken down into chemically meaningful components that correspond to an easily visualized picture of molecular structure. The approach also has some limitations. The potential problem is that MM routines will generate a conformation for which the strain energy is minimized. However, the minimum found during a calculation may not be the global minimum. It is relatively easy for the procedure to become trapped in a local energy minimum. There are schemes for minimizing the risk of such entrapment in local minima. For example, the calculation can be done a number of times starting from different initial geometries to see if the final geometry remains the same. Another obvious drawback of the MM approach is that it cannot be used to study any molecular system where electronic effects are dominant. Here, the QM approach, which explicitly accounts for the electrons in molecules, must be used.

To study electronic behavior in biomolecules, QM and MM are combined into one calculation (QM/MM) (Gogonea *et al.*, 2001; Warshel, 1991) that models a large molecule (e.g. enzyme) using MM and one crucial section of the molecule (e.g. active site) with QM. This is designed to give rapid results where only a particular region needs to be modeled quantum mechanically.

9.2.2 Energy calculation

A single point energy calculation is normally performed for a stationary point on a potential energy surface. The calculation provides energy and the gradient of that energy. The gradient is the root-mean-square (RMS) gradient of the derivative of the energy with respect to Cartesian coordinates, i.e.:

$$\text{RMS Gradient} = (3N)^{-1}[\Sigma(\delta E/\delta X)^2 + (\delta E/\delta Y)^2 + (\delta E/\delta Z)^2]^{1/2}$$

The local gradient in the potential energy defines the force field (FF). In MM calculations of the potential energy E_{total} (E_{total} is used here in accordance with most FF designation), the FFs generally take the form:

$$E_{total} = E_r + E_\theta + E\phi + E_{nb} + [\text{special terms}]$$

in which the successive terms, expressing the total energy (E_{total}), are energies associated with bond stretching (E_r), bond angle bending (E_θ), bond torsion ($E\phi$), nonbond interactions (E_{nb}) plus specific terms such as hydrogen bonding (E_{hb}) in biochemical systems. Most MM equations are similar in the terms they contain. There are some differences in the forms of the equations that can affect the choice of FF and parameters for the systems of interest. Examples are: i) MM2/3 for organic compounds (Altona and Faber, 1974) and ii) AMBER (Weiner *et al.*, 1984; Weiner *et al.*, 1986) or iii) CHARMm (Brooks *et al.*, 1983) for biological molecules.

For most FF, the internal energy terms are similar, namely

$$E_r = \Sigma K_r(r - r_0)^2$$
$$E_\theta = \Sigma K_\theta(\theta - \theta_0)^2$$
$$\text{and} \quad E\phi = \Sigma K_\phi[1 + \cos(n\phi - \phi_0)]$$

where K_r, K_θ and K_ϕ are force constants for bond, angle and dihedral angle respectively. Parameters, r_0, θ_0 and ϕ_0 define the equilibrium bond distance, equilibrium angle and phase angle for the given type. n is the periodicity of the Fourier term. These parameter values are derived from model molecules and vary between different FF.

The potential energy profile for the stretching of a chemical bond with a bond distance, r is an anharmonic function (steep ascending for $r < r_0$ and gentle ascending for $r > r_0$) with a minimum value at r_0. The MM treatment of the chemical bond approximates E_r and its profile is a harmonic function (symmetric ascending for $r < r_0$ and for $r > r_0$) without allowing an ultimate dissociation of the over-stretched chemical bond. This approximation depicts tighter bonding for molecules than would occur in reality. The potential energy function for deformation of the bond angle, θ is treated in a similar manner as that for the chemical bond, with θ_0 defining the equilibrium bond angle. However, the potential energy for twisting a dihedral angle, ϕ around the central bond (the bond between B and C of the four-atom center A-B-C-D) takes the form of the periodic function. For a single bond, the function defines three minima at $\phi = 0°$, 120° and 240°, associated with the staggered conformations around the bond. There are two minima at $\phi = 0°$ and 180° for a double bond. The planarity of aromatic, pyrimidine and purine rings may require explicit functions or stringent definitions of the bond angles and dihedral angles. The potential energies for bonding interactions, $E_b = E_r + E_\theta + E\phi$ are large but do not drive the folding of macromolecules because they are approximately the same for all conformations of the macromolecule. The conformations of macromolecules are defined by the weaker nonbonding interactions.

The nonbonding potentials, E_{nb} are derived from all the interactions that are not directly involved in covalent bonds. The potential energy functions for the nonbonding interactions are inversely related to some power of the distance between interacting atoms, $1/r^n$ (Chapter 1). Functions that depend on high powers of r (large n) are short-range interactions, whereas those with low powers of r (small n) are long-range interactions. In MM, different potential functions may be used by different FF for E_{nb} and special terms such as E_{hb} for hydrogen bonds. Most FF used in the modeling of biomacromolecules include E_{vdW} for van der Waals terms and E_{elec} for electrostatic terms in E_{nb}, for the interacting atoms *i* and *j*, with the distance of r_{ij}. Dipole–dipole interactions are generally incorporated into the potential function for electrostatic interactions by treating each atom as a monopole having a defined partial valence.

The Lennard–Jones 6–12 potential, which includes the first positive repulsive term and the second attractive term (London dispersion), is the most commonly used for van der Waals interactions, such that

$$E_{vdW} = \Sigma\Sigma(A_{ij}/r_{ij}^{12} - B_{ij}/r_{ij}^{6})$$

where A_{ij} and B_{ij} are van der Waal parameters. Together the repulsive potential and the attractive dispersion produce an equilibrium distance at which the two opposing energies become equal. The radii of neutral atoms that sum to give this equilibrium distance is known as van der Waal radii, r_{vdW}. The net force on the two atoms is zero at r_{vdW}. The steep increase in E_{vdW} at distances significantly shorter than the sum of r_{vdW} for two atoms is overcome by formation of a chemical bond.

Since biochemical molecules are often charged, an electrostatic energy term is added to E_{nb}:

$$E_{elec} = \Sigma\Sigma(q_i q_j)/(Dr_{ij})$$

in that D is a molecular dielectric constant (vary from 1 *in vacuo* to 80 in water) that account for the environmental attenuation of electrostatic interaction between the two atoms with the point charge q_i and q_j. The dielectric constant allows the MM function to account for the effects of a solvent on molecular structure without explicitly incorporating solvent molecules into the modeling. One strategy for assigning the dielectric constant to a macromolecule is to define a boundary that distinguishes the interior from the exterior of the macromolecule. For proteins in aqueous solutions, the exposed atoms can be set to that of water at the exterior of the protein. The interior of the protein is treated as a low dielectric cavity with typical values ranging from 1 to 20. Another strategy is to treat the dielectric constant as a distance-dependent variable.

The hydrogen bond (about 4–48 kJ/mol) is an interaction between a polarized D–H bond of a donor (D) and the polarized nonbonding orbitals of an acceptor (A). The bond is primarily a dipole–dipole interaction though, similar to covalent bond, there is an optimum distance separating the donor and acceptor atoms ranging from 0.26 to 0.30 nm. Importantly, this equilibrium distance is less than the sum of the respective r_{vdW} for the atoms involved. Thus the hydrogen bond is distinctively a special type of interaction that contributes significantly to the folding of biomacromolecules. The hydrogen bonding term, E_{hb} varies with different FF. Of the two most commonly used FF in biochemistry, AMBER introduces the 10–12 potential, i.e.:

$$E_{hb} = \Sigma\Sigma(C_{ij}/r_{ij}^{12} - D_{ij}/r_{ij}^{10})$$

while CHARMm considers both the distance and angle of the hydrogen bond interactions between three atoms (A for acceptor, H for hydrogen and D for donor).

One can choose to calculate all nonbonded interactions or to truncate (cut off) the nonbonded interaction calculations using a switched or shifted function. Useful guidelines for nonbonded interactions are:

- Calculate all nonbonded interactions for small and medium-sized molecules.
- Use either switched function or shifted function to decrease computing time for macromolecules, such as proteins and nucleic acids.
- Switched function is a smooth function, applied from the inner radius (R_{on}) to the outer radius (R_{off}), which gradually reduces a nonbonded interaction to zero. The suggested outer radius is approximately 14 Å and the inner radius is approximately 4 Å less than the outer radius.
- Shifted function is smooth function, applied over a whole nonbonded distance, from zero to outer radius that gradually reduces the nonbonded interaction to zero.

Thus different FF are designed for different systems and purposes. The databases used also differ. These differences must be borne in mind because a particular FF may work extremely well within one molecular structure class but may fail when applied to other types of structures.

9.2.3 Energy minimization

Molecular simulations include the application of FF to model the lowest potential energy of a conformation (energy minimization, EMin), the dynamic properties of macromolecular structure (molecular dynamics, MolD) and search for the optimum conformation of a macromolecule (conformation search). In all these techniques, the positions of atoms are perturbed in small increments, and it is difficult to sample all of the possible arrangements of atoms in conformational space. The successful simulation is the one that reproduces the experimentally observed properties of the molecule. It is essential to incorporate as much empirical information as possible into the initial model for simulation. While the overall system must be accurately defined, it is important to understand the limitations of FF since a choice of FF may affect the outcome of the simulation results.

The basic task in the computational portion of MM is to minimize the strain energy of the molecule by altering the atomic positions to optimal geometry. At an energy minimum, the gradient is zero. Thus the size of the gradient can provide qualitative information to assess if a structure is close to a minimum. This means minimizing the total nonlinear strain energy represented by the FF equation with respect to the independent variables, which are the Cartesian coordinates of the atoms (Altona and Faber, 1974). The following issues are related to the EMin of a molecular structure:

- The most stable configuration of a molecule can be found by minimizing its free energy, G.

- Typically, the energy E is minimized by assuming the entropy effect can be neglected.
- At a minimum of the potential energy surface, the net force on each atom vanishes, therefore the configuration is stable.
- Because the energy set at zero is arbitrary, the calculated energy is relative. It is meaningful only to compare energies calculated for different configurations of chemically identical systems.
- It is difficult to determine if a particular minimum is the global minimum, which is the lowest energy point where force is zero and second derivative matrix is definitely positive. Local minimum results from the net zero forces and positive definite second derivative matrix and saddle point results from the net zero forces and at least one negative eigenvalue of the second derivative matrix.

The most widely used methods fall into two general categories:

1. steepest descent and related methods such as conjugate gradient, which uses first derivatives, and
2. Newton–Raphson procedure, which additionally uses second derivatives.

The steepest descent method (Wiberg, 1965) depends on: i) either calculating or estimating the first derivative of the strain energy with respect to each coordinate of each atom; and ii) moving the atoms. The derivative is estimated for each coordinate of each atom by incrementally moving the atom and storing the resultant strain energy change. The atom is then returned to its original position, and the same calculation is repeated for the next atom. After all the atoms have been tested, their positions are all changed by a distance proportional to the derivative calculated in step (1). The entire cycle is then repeated. The calculation is terminated when the energy is reduced to an acceptable level. The main problem with the steepest descent method is that of determining the appropriate step size for atom movement during the derivative estimation steps and the atom movement steps. The sizes of these increments determine the efficiency of minimization and the quality of the result. An advantage of the first-derivative methods is the relative ease with which the force field can be changed.

The conjugate gradient method is a first-order minimization technique. It uses both the current gradient and the previous search direction to drive the minimization. Because the conjugated gradient method uses the minimization history to calculate the search direction and contains a scaling factor for determining step size, the method converges faster and makes the step sizes optimal as compared to the steepest descent technique. However, the number of computing cycles required for a conjugated gradient calculation is approximately proportional to the number of atoms (N) and the time per cycle is proportional to N^2. The Fletcher–Reeves approach chooses a descent direction to lower energy by considering the current gradient, its conjugate and the gradient for the previous step. The Polak–Ribiere algorithm improves on the Fletcher–Reeves approach by additional consideration of the previous conjugate and tends to converge more quickly.

The Newton–Raphson methods of EMin (Berkert and Allinger, 1982) use the curvature of the strain energy surface to locate minima. The computations are considerably more complex than the first-derivative methods, but they use the available information more fully and therefore converge more quickly. These methods involve setting up a system of simultaneous equations of size $(3N - 6)$ $(3N - 6)$ and solving for the atomic positions that are the solution of the system. Large matrices must be inverted as part of this approach.

The general strategy is to use steepest descents for the first 10–100 steps (500–1000 steps for proteins or nucleic acids), and then use conjugate gradients or Newton–Raphson to complete minimization for convergence (using RMS gradient or/and energy difference as an indicator). For most calculations, RMS gradient is set to 0.10 (values greater than 0.10 can be used for rapid, approximate calculations). The calculated minimum represents the potential energy closest to the starting structure of a molecule. The EMin is often used to generate a structure at a stationary point for a subsequent single point calculation or to remove excessive strain in a molecule, preparing it for a MolD simulation.

9.2.4 Molecular dynamics

Molecules are dynamic, undergoing vibrations and rotations continually. Therefore the static picture of molecular structure provided by MM is not realistic. Flexibility and motion are clearly important to the biological functioning of biomacromolecules. These molecules are not static structures, but exhibit a variety of complex motions both in solution and in the crystalline state. Energy minimization concerns only the potential energy term of the total energy and so it treats the biomacromolecule as a static entity. The dynamic properties of the atoms in a macromolecule or the momentum of the atoms in space requires the description of the kinetic term. The momentum (p) is related to the force exerted on the atom (Fi) and the potential energy (V) by

$$F = \partial p/\partial t = \partial V/\partial r.$$

Thus the kinetic energy is related to the potential energy functions of the FF. The kinetic energy of an atom is related to the temperature T of the system by a simplified form:

$$K = 3k_B T$$

where k_B is the Botzmann constant (1.38066×10^{-23} J/K), and the factor 3 accounts for the three directional components (x, y, z) for velocity.

The most commonly employed simulation method used to study the motion of biomacromolecules on the atomic level is the molecular dynamics (MolD) method (McCammon and Harvey, 1987). It is a simulation procedure consisting of the computation of the motion of atoms in a molecule according to Newton's laws of motion. The forces acting on the atoms, required to simulate their motions, are generally calculated using molecular mechanics FF. Rather than being confined to a single low-energy conformation, MolD allows the sampling of a thermally distributed range of intramolecular conformation. Molecular dynamics calculations provide information about possible conformations, thermodynamic properties and dynamic behavior of molecules according to Newtonian mechanics. A simulation first determines the force on each atom (F_i) as the function of time, equal to the negative gradient of the potential energy (V) with respect to the position (x_i) of atom *i*:

$$F_i = -\partial V/\partial x_i$$

The acceleration a_i of each atom is determined by:

$$a_i = F_i/m_i$$

The change in velocity v_i is equal to the integral of acceleration over time. One numerically and iteratively integrates the classical equations of motion for every explicit atom N in the system by marching forward in time via tiny time increments, Δt in MolD. A number of algorithms exist for this purpose (Brooks *et al.*, 1988; McCammon and Harvey, 1987) and the simplest formulation is shown below:

$$x_i(t + \Delta t) = x_i(t) + v_i(t)\Delta t$$
$$v_i(t + \Delta t) = v_i(t) + a_i(t)\Delta t = v_i(t) + \{F(x_1 \ldots x_N, t)/m\}\Delta t$$

The kinetic energy (K) is defined as

$$K = {}^1/_2\Sigma m_i v_i$$

The total energy of the system, called the Hamiltonian (H), is the sum of the kinetic (K) and potential (V) energies:

$$H(r, p) = K(p) + V(r)$$

where p is the momentum of the atoms and r is the set of Cartesian coordinates.

In MolD simulation of a biomacromolecule, the time interval, Δt and the average temperature <T> must be defined. From <T>, the initial velocities along the Cartesian axes, *x*, *y*, and *z* are set randomly to give a Gaussian distribution for the ensemble of atoms. This defines the kinetic energy of the system that along with the potential energy yields the total energy. The time increment must be sufficiently small that errors in integrating 6N equations (3N velocities and 3N positions) are kept manageably small, as manifested by conservation of the energy. As a result, Δt must be kept in the order of femtosecond (10^{-15} s) to picosecond (10^{-12}). Furthermore, because the forces, F must be recalculated for every time step, MolD is a computationally intensive task. Thus the overall time scale accessible to MolD calculations is in the order of picoseconds (10^{-12} s). One simplification to the system in MolD simulation is to reduce the number of atoms by not treating hydrogen atoms explicitly but uniting their properties with those of the attached heavy atom, e.g. treating C–H as a single mass. The molecular simulation approximates the condition in which the total energy of the system does not change during the equilibrium simulation. One way to test for the success of a dynamics simulation and the length of the time step is to determine the change in kinetic and potential energy between time steps. In the microcanonical ensemble (constant number, volume and energy), the change in kinetic energy should be of the opposite sign and exact magnitude as the potential energy.

The MolD normally consists of three phases; heating, equilibration and cooling. To perform MolD, the structure is submitted to a minimization procedure to relieve any strain inherent in the starting positions of the atoms. The next step is to assign velocities to all the atoms. These velocities are drawn from a low-temperature Maxwellian distribution. The system is then equilibrated by integrating the equations of motion while slowly raising the temperature and adjusting the density. The temperature is raised by increasing the velocities of all of atoms. There is a simple analytical function expressing the relationship between kinetic energies of the atoms and the temperature of the system:

$$T(t) = 1/\{k_B(3N - n)\}\sum_{i=1}^{N} m_i|v_i|$$

where

$T(t)$ = temperature of the system at time t

$(3N - n)$ = number of degrees of freedom in the system

v_i = velocity of atom i at time t

k_B = Boltzmann constant

m_i = mass of atom i

N = number of atoms in the system

This process of raising the temperature of the system will cover a time interval of 10–50 ps. The period of heating to the temperature of interest is followed by a period of equilibration with no temperature changes. The stabilization period will cover another time interval of 10–50 ps. The mean kinetic energy of the system is monitored, and when it remains constant, the system is ready for study. The structure is in an equilibrium state at the desired temperature.

The MolD experiment consists of allowing the molecular system to run free for a period of time, saving all the information about the atomic positions, velocities, and other variables as a function of time. This (voluminous) set of data is called a trajectory. The length of time that can be saved during trajectory sampling is limited only by the computer time available and the speed of the computer. Once a trajectory has been calculated, all the equilibrium and dynamic properties of the system can be calculated from it. Equilibrium properties are obtained by averaging over the property during the time of the trajectory. Plots of the atomic positions as a function of time schematically depict the degree to which molecules are moving during the trajectory. The RMS fluctuations of all of the atoms in a molecule can be plotted against time to summarize the aggregate degree of fluctuation for the entire structure. The methods of MolD are becoming an important component for the study of biomacromolecular structures in an effort to rationalize structural basis for their activities and functions.

Molecular dynamics simulations are efficient for searching the conformational space of medium-sized molecules, especially ligands in free and complexed states. Quenched dynamics is a combination of high temperature MolD and EMin. For a conformation in a relatively deep local minimum, a room temperature MolD simulation may not overcome the barrier. To overcome barriers, conformational searches use elevated temperature (>600 K) at constant energy. To search conformational space adequately, simulations are run for 0.5–1.0 ps each at high temperature. For a better estimate of conformations, the quenched dynamics should be combined with simulated annealing, which is a cooling simulation. Cooling a molecular system after heating or equilibration can:

- reduce stress on molecules caused by a simulation at elevated temperatures and take high energy conformational states toward stable conformations; and
- overcome potential energy barriers and force a molecule into a lower energy conformation.

Quenched dynamics can trap structures in a local minimum. The molecular system is heated to elevated temperatures to overcome potential energy barriers and then cooled slowly to room temperature. If each structure occurs many times during the search, one is assured that the potential energy surface of that region has been adequately sought.

The molecular dynamics is useful for calculating the time dependent properties of an isolated molecule. However, molecules in solution undergo collisions with other molecules and experience frictional forces as they move through the solvent. Langevin dynamics simulates the effect of molecular collisions and the resulting dissipation of energy that occur in real solvents without explicitly including solvent molecules by adding a frictional force (to model dissipative losses) and a random force (to model the effect of collisions) according to the Langevin equation of motion:

$$a_i = F_i/m_i - \gamma v_i + R_i/m_i$$

where γ is the friction coefficient of the solvent and R_i is the random force imparted to the solute atom by the solvent. The friction coefficient determines the strength of the viscous drag felt by atoms as they move through the medium and γ is the friction coefficient related to the diffusion constant (D) of the solvent by $\gamma = k_BT/mD$. At low values of the friction coefficient, the dynamic aspects dominate and Newtonian mechanics is recovered as $\gamma \rightarrow 0$. At high values of γ, the random collisions dominate and the motion is diffusion-like.

Monte Carlo (MC) simulations (Allen and Tildesley, 1967) can be used to conduct conformational searches under nonequilibrium conditions. Unlike MolD or Langevin dynamics, which calculate ensemble averages by calculating averages over time, MC calculations evaluate ensemble averages directly by sampling configurations from the statistical ensemble. To generate trajectories that sample commonly occurring configurations, the Metropolis method (Metropolis *et al.*, 1953) is generally employed. Thermodynamically, the probability of finding a system in a state with ΔE above the ground state is proportional to $\exp(-\Delta E/kT)$. Thus if the energy change associated with the random movement of atoms is negative, the move is accepted. If the energy change is positive, the move is accepted with probability $\exp(-\Delta E/kT)$.

9.2.5 Conformational search

Conformational search is a process of finding low energy conformations of molecular systems by varying user-specified dihedral angles. The method involves variation of dihedral angles to generate new structures and then energy minimizing each of these angles. Low energy unique conformations are stored while high-energy duplicate structures are discarded. Because molecular flexibility is usually due to rotation of unhindered bond dihedral with little change in bond lengths or bond angles, only dihedral angles are considered in the conformational search. Its goal is to determine the global minimum of the potential energy surface of a molecular system. Several approaches have been applied to the problem of determining low energy conformations of molecules (Howard and Kollman, 1988). These approaches generally consist of the following steps with differences in details:

1. *Selection of an initial structure*: The initial structure is the most recently accepted conformation (e.g. energy minimized structure) and remains unchanged during the search. This is often referred to as a random walk scheme in MC searches. It is based on the observation that low energy conformations tend to be similar, therefore starting from an accepted conformation tends to keep the search in a low energy region of the potential surface. An alternative method, called the usage directed method, seeks to uniformly sample low energy region by going through all previously accepted conformations while selecting each initial structures (Chang *et al.*, 1989). Comparative studies have found the usage directed scheme to be superior in quickly finding low energy conformations.
2. *Modification of the initial structure by varying geometric parameters*: The variations can be either systematic or random. Systematic variations can search the conformational space exhaustively for low energy conformations. However, the number of variations becomes prohibitive except for the simplest systems. One approach to reduce the dimensionality of systematic variation is to first exhaust variations at a low resolution, then exhaust the new variations allowed by successively doubling the resolution. Random variations choose a new value for one or more geometric parameters from a continuous range or from sets of discrete values. To reduce the number of recurring conformations, several random variations have some sort of quick comparison with the sets of previous structures prior to performing EMin of the new structure.
3. *Geometry optimization of the modified structure to energy-minimized conformations*: The structures generated by variations in dihedral angles are energy minimized to

find a local minimum on the potential surface. Although the choice of optimizer (minimizer) has a minor effect on the conformational search, it is preferable to employ an optimizer that converges quickly to a local minimum without crossing barrier on the potential surface.

4. *Comparison of the conformation with those found previously*: The conformation is accepted if it is unique and its energy satisfies a criterion. Two types of criteria are used to decide an acceptance of the conformation. First, geometric comparisons are made with previously accepted conformations to avoid duplication. Conformations are often compared by the maximum deviation of torsions or RMS deviation for internal coordinates, interatomic distances or least-squares superposition of conformers. Because geometry optimization can invert chiral centers, the chiral centers of the modified structures should be checked after EMin. Second, the energetic test for accepting a new conformer may be carried out by a simple cutoff relative to the best energy found so far or a Metropolis criterion where higher energy structures are accepted with a probability determined by the energy difference and a temperature, e.g. $\exp(-\Delta E/kT)$.

9.2.6 Remaining issues

The molecular mechanical approach to simulating biomacromolecular structures by the use of potential energy functions has been discussed. These potential energy functions are an enthalpic contribution to the free energy, but the free energy contains entropic contribution. There are two forms of entropy in a biomacromolecular system; namely the conformational entropy, which entails the inherent entropy of the biomacromolecular structure and the solvent entropy, which results from interactions between a biomacromolecule and solvent molecules.

The conformational entropy of a macromolecule measures its degree of conformational freedom. It is related to the degree of degeneracy at each energy level of the biomacromolecule and can be estimated from the degree of rotational freedom about the freely rotating bonds of the chain. For a biomacromolecule consisting of n monomers, the number of possible conformations, N can be estimated from the number of possible conformations, g for each monomeric unit by

$$N = g^{(n-2)}$$

Thus the difference in entropy between the native and all possible random structures is

$$\Delta S = S_N - S_R = k_B\{\ln(1) - \ln N\} = k_B(n-2)\ln g$$

An approach to estimating conformational entropy is to calculate the density of the folded state (ρ_F) versus that of the unfolded state (ρ_U), because the density reflects the effective volume occupied by each chain of a biomacromolecule with a well-defined mass. At each step of folding, the macromolecular chain samples successively higher density states with smaller effective volume. Thus folding of a biomacromolecule to a compact structure reduces the volume and lowers the entropy. The difference in entropy for any two states with the densities, ρ_1 and ρ_2 relative to the density of the native state, ρ_0 is

$$\Delta S = -3k_B/2[(\rho_0/\rho_2)^{2/3} - (\rho_0/\rho_1)^{2/3}]$$

In this evaluation of conformational entropy, the isomers and the degeneracy are not counted and ΔS is dependent only on the relative densities of each state along the folding pathway. The expression suggests a successive decrease in the entropy during which a less compact structure of a biomacromolecule is folded to more compact structures, though the number of possible conformations decreases, as the structure becomes more compact.

The solvent entropy is considered to be the thermodynamic contribution of the hydrophobic effect, which is primarily the driving force that folds biomacromolecules into compact structures. Hydration/solvation is an important problem in simulation of biomacromolecules since all biomacromolecules interact with water/aqueous solvents and the sequestering of nonpolar groups away from water gives rise to the hydrophobic effect. A simple relationship to describe the free energy for hydrating/solvating a macromolecule is

$$\Delta G^{o}_{H} = \Delta G^{o}_{vdW} + \Delta G^{o}_{cav} + \Delta G^{o}_{e}$$

where ΔG^{o}_{H}, ΔG^{o}_{vdW}, ΔG^{o}_{cav}, and ΔG^{o}_{e} are free energy of hydration/solvation, favorable van der Waals interactions of the first-shell water molecules with the atoms of the macromolecule, unfavorable entropic formation of a clatherate cavity in the solvent to fit the macromolecule and other electrostatic or dipolar interactions between the macromolecule and the solvent dipoles respectively (Sharp *et al.*, 1991).

The remaining task involves putting together enthalpic contributions from the potential energy function and entropic contributions, to compute the free energy of the system, which is the measure of biomacromolecular stability. Experimentally, the free energy (the standard free energy ΔG^{o} at equilibrium) for the system is determined by $\Delta G = \Delta G_{N} - \Delta G_{U}$, in which ΔG_{N} and ΔG_{U} are free energies for the folded and unfolded structures respectively. The free energy can be evaluated using the method that samples conformation space. The conformations are generated and evaluated by MolD simulation to give time-average free energy change ΔG between free energies G and G* for two states with potentials V(Γ) and V*(Γ) respectively, according to

$$\Delta G = -k_{B}T\ln<e^{-V^{*}(\Gamma)-V(\Gamma)}>$$

where Γ represents one point in conformation space with V(Γ) as its potential for that conformation. The term, $<e^{-V^{*}(\Gamma)-V(\Gamma)}>$ describes an average over a Boltzmann population for conformations sampled in the system corresponding to the sum of $e^{-\Delta V(\Gamma)/k}B^{T}$ for all conformations divided by the total number of conformations sampled.

MC simulations are commonly used to compute the average thermodynamic properties of a molecule or a molecular system, especially the structure and equilibrium properties of liquids and solutions (Allen and Tildesley, 1967). In this method, an arbitrary starting conformation Γ is generated and its energy is calculated. The structure is perturbed slightly to a new conformation Γ* and the energy recalculated. If $V^{*}(\Gamma) < V(\Gamma)$, the new conformation is kept, whereas the new conformation is either kept or discarded randomly if $V^{*}(\Gamma) > V(\Gamma)$ by comparing $\exp\{-\Delta V(\Gamma)/k_{B}T\}$ to a random number i_{R} that varies between 0 and 1. The macromolecule is allowed to sample sufficiently large areas of conformation space to accurately describe the average state of the system. MC simulations lead toward a minimum free energy and not simply to the lowest energy of the potential surface of the system.

9.2.7 Computational application of molecular modeling packages

The facile conversion of sequences into 3D structures that can be displayed and manipulated on the computer screen (molecular graphics) has greatly improved molecular modeling as an essential tool in biochemical research and teaching. The 1D sequences can be converted into 2D structural representations by the use of ISIS Draw, which can be down-

TABLE 9.1 Molecular modeling (MM) programs of biochemical interest

MM/FF program	Source	Reference
AMBER	Univ of Calif, San Francisco/HyperChem	Weiner *et al*., 1984
CHARMM	Harvard Univ/Accelrys, Inc.	Brooks *et al*., 1983
ECEPP	Cornell Univ	Nemathy *et al*., 1983
GROMOS	Univ of Groningen/Biomos	Herman *et al*., 1984
SYBYL	Tripos, Inc.	Clark *et al*., 1989
MM2/3	Univ of Georgia/Chem3D	Allinger *et al*., 1989
MACROMODEL	Columbia Univ	Chemistry, Columbia University
OPLS	Yale Univ/HyperChem	Jorgensen and Tirado-Rives, 1988

Note: References are: Allinger *et al.* (1989); Brooks *et al.* (1983); Clark *et al.* (1989); Herman *et al.* and (1984); Jorgensen and Tirado-Rives (1988); Nemathy *et al.* (1983); Weiner *et al.* (1984).

loaded from the MDL Information system (http://www.mdli.com/dwonload/isisdraw.html) or numerous commercial software modeling packages, which normally combine 2D representation, 3D visualization and modeling. Molecular graphics programs that are widely used to display biomolecular structures include RasMol (http://www.umass.edu/microbio/rasmol/) KineMage (http://orca.st.usm.edu/~rbateman/kinemage/) and Cn3D (http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.html).

Published FF parameters for MM (Jalaie and Lipkowitz, 2000; Osawa and Lipkowitz, 1995) and software for molecular modeling (Boyd, 1995) have been compiled. Some of the MM programs applicable to molecular modeling of biomolecules are listed in Table 9.1.

All of these programs are available for Unix operating system. Most of the Windows versions are incorporated into commercial molecular modeling packages, such as MM in Chem3D of CambridgeSoft (http://www.camsoft.com), AMBER and CHARMm in HyperChem of HyperCube (http://www.hyper.com), SYBYL in PC Spartan of WaveFunction (http://www.wavefun.com). Online AMBER server (http://www.amber.ucsf.edu/amber/amber.html) conducts *in vacuo* minimization with AMBER 5.0 and then electrostatic solvation with AMBER 6.0. B server (http://www.scripps.edu/~nwhite/B/indexFrames.html) implements AMBER, and Swiss-Pdb Viewer (http://www.expasy.ch/spdbv/mainpage.html) executes GROMOS. Some aspects of molecular modeling using HyperChem are shown in Table 9.2.

9.3 STATISTICAL THERMODYNAMICS

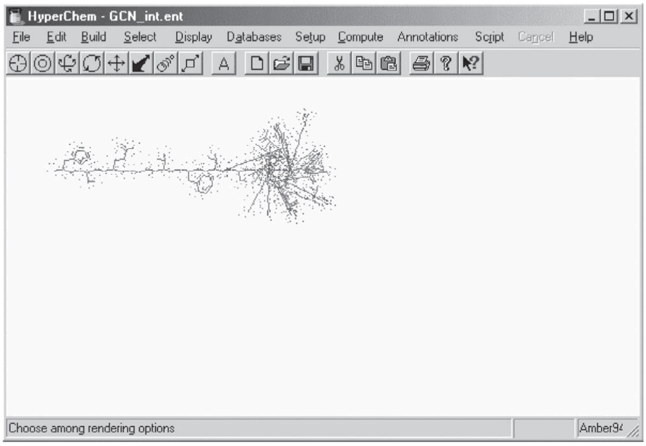
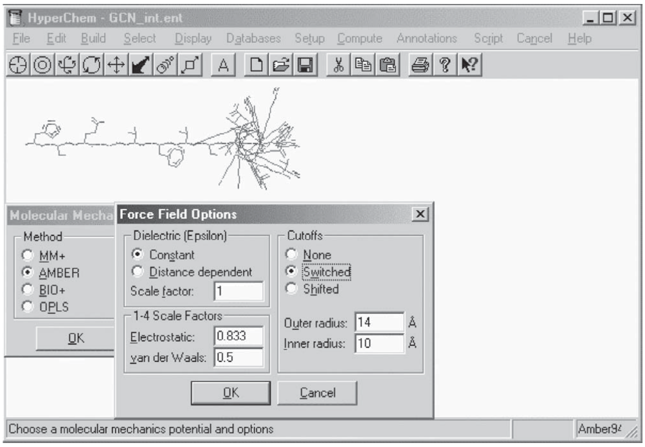
9.3.1 General principles

Biomacromolecular properties observed in solution and in cells reflect the average behavior of a population. Statistical mechanics allows a description of the distribution of molecular conformations that contribute to this population under a given thermodynamic state. The average properties of a biomacromolecule can be represented as either the time-average behavior of a single molecule or the behavior of an ensemble of many molecules at any instant in time. A basic postulate of statistical mechanics is that the time average of a certain property of a system is the same as the ensemble average at any instant. Therefore according to the statistical mechanical treatment of thermodynamic properties known as statistical thermodynamics, the time-average and ensemble average behaviors are identical. A time-average conformation of a macromolecular structure can be modeled by per-

TABLE 9.2 Example for energy minimization (geometric optimization). Molecular mechanics energy minimization with Amber is exemplified for geometric optimization of glucagon structure using HyperChem

Modeling	Result/Visualization

Sequence retrieval and prediction of secondary structures
1. Retrieve amino acid sequence from Entrez in fasta format
2. Submit the sequence to NPS@ for consensus secondary structure prediction choosing Gor4, PHD and others as prediction methods
	`Glucagon     HSQGTFTSDYSKYLDSRRAQDFVQYLMNT`
	`DSC          ccccccccchhhhhhhhhhhhhhhhhccc`
	`GOR4         ccccceeecchhhhhhhhhhhheeeeeeec`
	`HNNC         ccccccccchchcccchhhhhhhhhhhcc`
	`PHD          cccccccchhhhhchhhhhhhhhhhhhcc`
	`SIMPA96      cccccchhhhhhhhhhhchhhhhhhhhhc`
	`Sec. Cons.` **`ccccccccchhhhhhhhhhhhhhhhhhcc`**

Construction of initial model based on consensus secondary structure: 1. Build model of glucagon based on predicted secondary structure, i.e. coils for residues 1–9 and 28–29, and α-helix for residues 10–27 via Databases → Amino acids 2. Select conformation, others for H1–D9, helix for Y10–M27 and others for N28–T29 3. Display → Sticks and dots as shown 4. Save the file	 HyperChem - GCN_int.ent File Edit Build Select Display Databases Setup Compute Annotations Script Cancel Help Choose among rendering options Amber94
Set up molecular mechanics calculation: 1. Choose EFF, i.e. Amber *via* Setup → Molecular mechanics → Amber 2. Choose options: e.g. dielectric constant, scale factors and cutoff condition	 HyperChem - GCN_int.ent File Edit Build Select Display Databases Setup Compute Annotations Script Cancel Help Molecular Mecha Force Field Options OK OK Cancel Method: ○ MM+ ● AMBER ○ BIO+ ○ OPLS Dielectric (Epsilon): ● Constant ○ Distance dependent Scale factor: 1 1-4 Scale Factors Electrostatic: 0.833 van der Waals: 0.5 Cutoffs: ○ None ● Switched ○ Shifted Outer radius: 14 Å Inner radius: 10 Å Choose a molecular mechanics potential and options Amber94

(continued)

TABLE 9.2 ***continued***

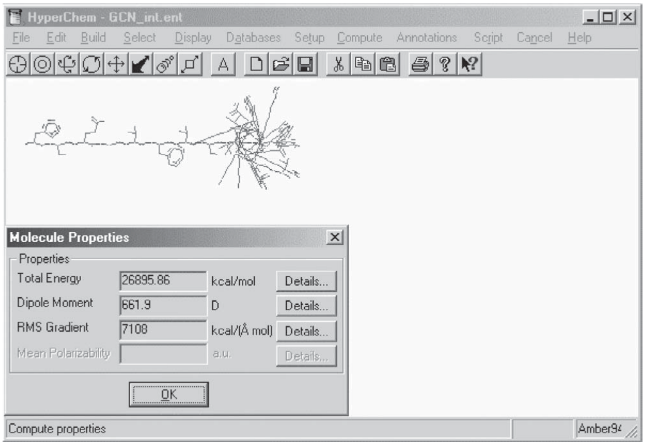
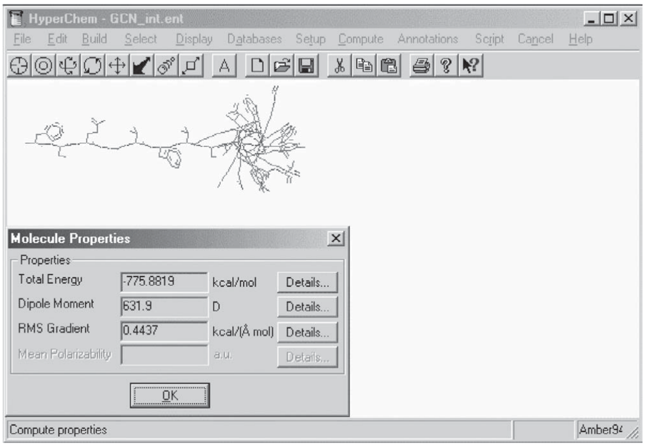
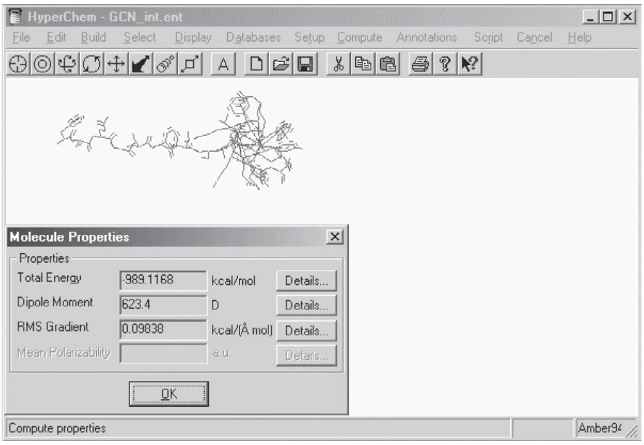
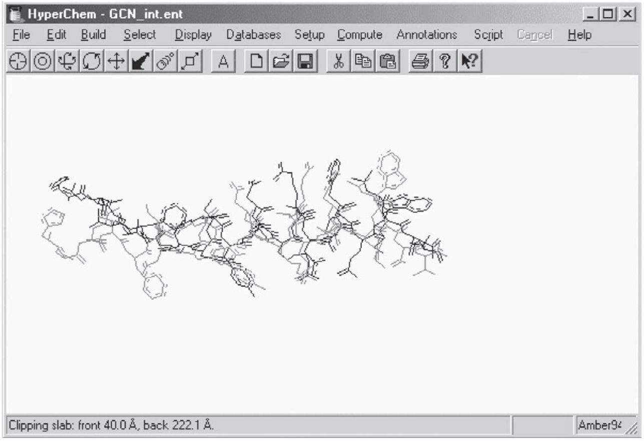
Modeling	Result/Visualization
Single point energy calculation: 1. Compute → single point energy 2. Compute → properties. The initial model shows total energy (E_t) of 29895.86 kcal/mole and RMS gradient (RMSG) of 7108 kcal/Å-mole	 HyperChem - GCN_int.ent File Edit Build Select Display Databases Setup Compute Annotations Script Cancel Help Molecule Properties Properties Details... OK Amber94 Total Energy 26895.86 kcal/mol Dipole Moment 661.9 D RMS Gradient 7108 kcal/(Å mol) Mean Polarizability a.u. Compute properties
Geometric optimization with steepest descent (SD): 1. Setup optimization condition by Compute → geometric optimization 2. First choose Steepest descent as the algorithm with RMSG = 0.1 kcal/Å-mole or maxi. cycles = 1000 as the termination condition 3. The SD optimization results in total energy = −775.89 kcal/mole and RMSG = 0.4437 kcal/Å-mole	 HyperChem - GCN_int.ent File Edit Build Select Display Databases Setup Compute Annotations Script Cancel Help Molecule Properties Properties Details... OK Amber94 Total Energy -775.8819 kcal/mol Dipole Moment 631.9 D RMS Gradient 0.4437 kcal/(Å mol) Mean Polarizability a.u. Compute properties
Geometric optimization with conjugate gradient (CG): 4. Next choose Polak-Ribiere (conjugate gradient) as the algorithm with RMSG = 0.1 kcal/Å-mole or maxi. cycles = 1000 as the termination condition twice 5. The CG optimization terminates after 1055 cycles results in E_t = −989.12 kcal/mole and RMSG = 0.09838 kcal/Å-mole	 HyperChem - GCN_int.ent File Edit Build Select Display Databases Setup Compute Annotations Script Cancel Help Molecule Properties Properties Details... OK Amber94 Total Energy -989.1168 kcal/mol Dipole Moment 623.4 D RMS Gradient 0.09838 kcal/(Å mol) Mean Polarizability a.u. Compute properties

TABLE 9.2 *continued*

Modeling	Result/Visualization
Comparison with known 3D structure of glucagon: 1. Retrieve x-ray structure of porcine glucagon (1GCN.pdb) from PDB 2. Display the optimized model structure 3. File → Merge to display 1GCN 4. Select the model and color it black 5. Select 1GCN and color it red 6. Select residues in the helical region for overlapping *via* Display → Residues → Number ≥10 and ≤27 7. Display → RMS Fit and Overlay	 HyperChem - GCN_int.ent File Edit Build Select Display Databases Setup Compute Annotations Script Cancel Help A Clipping slab: front 40.0 Å, back 222.1 Å. Amber94

Note: Web sites used are: Enztrez at http://www.ncbi.nlm.nih.gov/Entrez, NPS@ at http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa-server.html and PDB at http://www.rcsb.org/pdb.

forming MolD simulation while information on the distribution of macromolecular conformations can be obtained by statistical methods. The statistical concepts provide a connection between the molecular dynamical states of molecules and the conventional thermodynamic laws for bulk systems. These are useful in conformational analysis, structural transition, molecular interaction/binding and association-dissociation studies of biomacromolecules. Indeed, thermodynamic properties of a macroscopic system can be obtained from the molecular parameters of its microscopic constituents (Eriksson, 2001; Sun, 1994; Van Holde *et al.*, 1998). In this section, principles of statistical mechanics to study multiple thermodynamic states of biomacromolecules, will be introduced.

The simplest multiple state system involves molecules that can exist in one of two states, A and B, in which A serves as the reference state to which B is compared. At equilibrium:

$$A \leftrightarrow B$$
$$K_{eq} = [B]/[A]$$

Thermodynamic parameters, standard free energy change (ΔG^o), standard enthalpy change (ΔH^o) and standard entropy change (ΔS^o) may be obtained from the usual relationships:

$$\Delta G^o = -RT \ln K_{eq},$$
$$\Delta H^o = -R[\partial(\ln K_{eq})/\partial(1/T)]_p$$

and

$$\Delta S^o = (\Delta H^o - \Delta G^o)/T$$

in which R is the gas constant ($8.3144\,J^oK^{-1}\,mol^{-1}$).

The equilibrium constant defines the probability of observing the molecule as B relative to the probability of observing A. Thus the probability of finding B, P_B against all possible states, in the two-state example is

$$P_B = [B]/([A] + [B]) = K_{eq}/(1 + K_{eq})$$

An introduction of the statistical weight, w as the concentration of all species at some state, e.g. B relative to that of reference state, i.e. A such that $w_A = [A]/[A] = 1$ and $w_B = [B]/[A]$:

$$P_B = w_B/(1 + w_B)$$

For any state, j at equilibrium, it can be written as

$$G^o_j - G^o_0 = \Delta G^o_j = -RT \ln w_j$$

or

$$w_j = \exp[-\Delta G^o_j/RT]$$

where G^o_j, and G^o_0 are free energies for any state *j* and the reference state.

In this simple example, each state represents a single conformation of the molecule, A or B. However, if a particular state represents more than one isoenergetic conformation of the molecule, it is degenerate and a higher statistical probability of that state is observed. The degeneracy of each state *j* is reflected in the intrinsic entropy of that state, i.e. $S_j = R \ln g_j$ and is incorporated into the partition function, Q as

$$Q = \sum_{j=0}^{N} g_j w_j$$

The parameter g_j is called the degeneracy of the state *j*. Thus the probability P_j of observing a particular state *j* of a system becomes

$$P_j = g_j w_j/Q$$

For the treatment of chain conformation problems in biomacromolecules such as two-state structural transitions to be discussed in the next section, a partition function can be constructed as follows:

1. Consider all possible conformations of the chain.
2. Establish a reference state for a chain element, and assign this state a statistical weight of unity.
3. For each conformation of the chain, assign an appropriate statistical weight for each element that is not in the reference state.
4. Construct the partition function by summing over all conformations, with appropriate products of statistical-weighting factors assign to each conformation.

9.3.2 Transitions of regular structures: Two-state models

The probability of observing one of two possible states can be demonstrated with a statistical mechanics treatment of transitions between regular structures of biomacromolecules. Examples include transitions between α-helix and β-strand or α-helix and random coil in proteins, between double- and single-stranded structures in nucleic acids, and between helical and coil conformations in glycans. To assign chain elements, we can initially assume that the reference state A has all residues of the polymer chain in the *a* conformation and the new state B have all residues in the *b* conformation. Each state will be referred to by the overall statistical weights, w_0 for the reference state and w_j, including the degeneracy for each state *j*. If N number of residues simultaneously converts from *a* to *b*, the molecule consists of two possible states with statistical weights w_0 and w_N; this is an all-or-none transition (all residues are either *a* or *b*). The transition of each residue

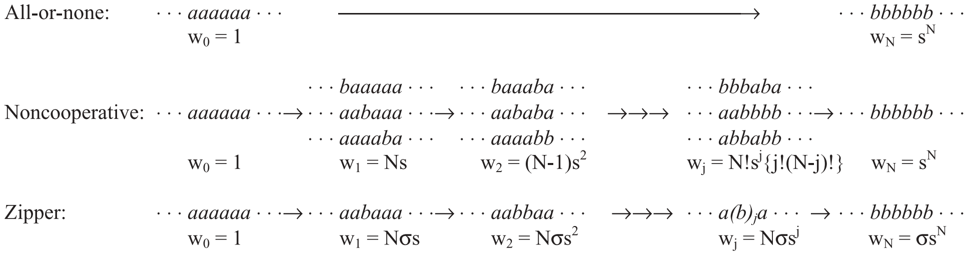

All-or-none:
· · · aaaaaa · · ·
w0 = 1
· · · bbbbbb · · ·
wN = s^N
Noncooperative:
· · · aaaaaa · · ·
· · · baaaaa · · ·
· · · aabaaa · · ·
· · · aaaaba · · ·
· · · baaaba · · ·
· · · aababa · · ·
· · · aaaabb · · ·
· · · bbbaba · · ·
· · · aabbbb · · ·
· · · abbabb · · ·
· · · bbbbbb · · ·
w0 = 1
w1 = Ns
w2 = (N-1)s^2
wj = N!s^j{j!(N-j)!}
wN = s^N
Zipper:
· · · aaaaaa · · ·
· · · aabaaa · · ·
· · · aabbaa · · ·
· · · a(b)ja · · ·
· · · bbbbbb · · ·
w0 = 1
w1 = Nσs
w2 = Nσs^2
wj = Nσs^j
wN = σs^N

Figure 9.1 Models and statistical weights for structural transitions

within the biomacromolecule defines a statistical weight for each residue, s ([*b*]/[*a*]) that expresses the probability for a residue to convert from *a* to *b* (Figure 9.1). Thus the statistical weight of the overall state w_N is the product of s for each residue in a biomacromolecule of N residues, i.e.:

$$W_N = \prod_{j=1}^{N} s_j = S^N \qquad \text{since all } s_j \text{ are identical and equal to s.}$$

All-or-none transition is highly cooperative with N being the cooperative length of a transition.

In the noncooperative model (Figure 9.1), the biomacromolecule converts stepwise from *a* to *b*, with an increase in the fraction of *b* with each transition. Each step *j* represents a different state *j* with w_j being the statistical weight in which *j* also represents the total number of residues in *b* conformation with the probability, s = [*b*]/[*a*]. There are N unique combinations, i.e. degeneracy with a single residue *b* for a chain of N residues, therefore $w_1 = Ns$ and $w_j = g_j s^j$. The general form of the partition function for N number of residues is the polynomial expression:

$$Q = (1 + s)^N$$

The binomial expansion gives

$$(1 + s)^N = 1 + Ns + [N \cdot (N - 1)/2!]s^2 + \ldots + [\{N!/\{j!(N - j)!\}]s^j + \ldots s^N$$

remembering $N! = N \cdot (N - 1) \cdot (N - 2) \ldots 3 \cdot 2 \cdot 1$.

The expansion of this expression yields the binomial coefficients of the power of s (the degeneracy factors):

$$g_j = N!/\{j!(N - j)!\}$$

This is the number of ways in which to divide N residues into *j* in state *b* and N − *j* in state *a*, thus

$$W_j = g_j s^j = N!s^j/\{j!(N - j)!\}$$

The probability of observing any state *j* average across all states $<P_j>$ is

$$\langle P_j \rangle = \frac{W_j}{Q} = \frac{N!s^j/\{j!(N-j)!\}}{1 + \Sigma N!s^j/\{j!(N-j)!\}}$$

The fraction in *b* at each state is the number of *b* residues, *j* relative to the total number of residues, i.e. $f_j = j/N$. The total fraction of residues as *b* in the system is the average probability of observing *b*, $<P_b>$ which is the sum of all the states having at least one residue *b* conformation, with each state weighted by f_j, relative to all the possible states of the system. This is given by

$$\langle P_b \rangle = \frac{\Sigma(j/N)w_j}{Q} = s/(s+1)$$

The probability of observing *a* conformation becomes $<P_a> = 1 - <P_b>$, i.e.

$$\langle P_a \rangle = \frac{\Sigma\{(N-1)/N\}w_j}{Q} = 1/(s+1)$$

and as expected, the ratio, $<P_b>/<P_a> = s$.

Most transitions in the secondary structures of biomacromolecules fall somewhere between the cooperative none-or-all and the noncooperative models. Many of these transitions can be described by the zipper model, which dissects the structural transition of a polymeric chain into a number of discrete steps (Figure 9.1). The model is a special case of the cooperative structural transition of biomacromolecules. In the zipper model, the initiation of the transition is harder than extension (propagation) and therefore low probability. This initiation step is of high energy and provides a nucleation point for the transition. The subsequent extension steps occur by a series of lower energy and consequently higher probability.

The partition function for the zipper model is derived from the basic relationships of the noncooperative model. The only difference is the statistical weight for the first step w_1 that must include a nucleation parameter, σ to represent the probability (lower probability therefore $\sigma < 1$) for initiating the transition. Therefore the statistical weight for the state $j = 1$ is $w_1 = \sigma s$. Two possibilities exist for the next step of the transition. In one case, the next transition simply extends *b* to an immediately adjacent residue (e.g. . . . *aaabbaaa* . . .) with $w_2 = \sigma s^2$. An alternative case involves the conversion at a nonadjacent residue (e.g. . . . *aaabaaab* . . .) with $w_2 = \sigma^2 s^2$. The probability of observing two isolated residues versus two adjacent residues as *b* is $\sigma^2 s^2/\sigma s^2 = \sigma$. If $\sigma \ll 1$, the probability of having two isolated residues as *b* is extremely low in the structural transitions. In the zipper model with adjacent extension of *b*, each subsequent step along this contiguous region introduces additional s in w_j for each state *j* in

$$w_j = \sigma s^j$$

The term s is referred to as the propagation parameter since it describes the statistical weight for propagating the transition by one residue. If only the contiguous state is considered being in either *a* or *b*, the degeneracy for any state *j* for $j = 1$ to N, the number of ways of fitting a segment of *j* adjacent residues into N possible position is

$$g_j = N - j + 1$$

Thus the partition function for the zipper model becomes

$$Q = 1 + \sum_{j=1}^{N} w_j = 1 + \sum_{j=1}^{N} (N - j + 1)\sigma s^j = 1 + \sigma \sum_{j=1}^{N} (N - j + 1)s^j$$

The average probability of observing *b* in the distribution of macromolecule, $<P_b>$ for $j \geq 1$ is:

$$\langle P_b \rangle = \frac{\sigma\Sigma(f_j w_j)}{Q} = \frac{\sigma\Sigma[(j/N)(N - j + 1)s^j]}{Q}$$

The cooperativity of the structural transition via the zipper model is dependent on the magnitude of σ relative to s. If $\sigma = 1$, the model does not discriminate between the first and subsequent transitions to *b*, and thus the transition is noncooperative. On the other hand, if $\sigma \ll$ s, the transition is cooperative and approaches the all-or-none model. Because the cooperativity of the structural transition is dependent on σ, it is termed the cooperativity coefficient. The zipper model can describe, in addition to structural transition, a large variety of processes involving biomacromolecules such as the formation of regular structures along polymeric chains, association-dissociation of multimeric complexes, structural perturbations induced by interactions with ligands and changes in external environments. The nucleation parameter, σ is typically intrinsic to the type of transition being studied, thus σ determined for a specific transition process of a given chain length should be applicable to the same process for polymeric chains of any length. The propagation parameter, *s* is dependent on the factors that stabilize/destabilize *b* versus *a* with respect to each residue along the polymeric chain and can be described by

$$s = \exp[-(\Delta G^o_I + \delta)/RT]$$

where ΔG^o_I is the intrinsic energy difference between *b* and *a*, and δ is an externally induced perturbation to the system. The dependence of *s* on external factors provides the impetus for a shift in the macromolecular population from the reference state to some new state.

9.3.3 Random structure: Random-walk problem

The statistical methods can also be used to model and study nonregular (random) structures, including tertiary structures of biomacromolecules. A random structure is one in which the conformation of each monomer unit along a chain is entirely independent of all other units. The statistical problem of treating random structures is analogous to a phenomenon of random walk involving someone who walks along a sidewalk in random directions unable to decide at each step whether to go forward or backward. Although the accurate description of random conformation for a biomacromolecule is in 3D, many of the properties for the simple one-dimensional random walk also apply to 3D and therefore the one-dimensional random walk will be exemplified.

In a one-dimensional random walk, allowing each step to move forward or backward along a line from the origin, let *f* be the probability or statistical weight of a forward step and *b* the probability (statistical weight) of backward step. Since this is random, the probability of stepping forward and that of stepping backward are equal, i.e. $f = b$ and $f + b = 1$. For N total steps, there are *j* steps forward and $N - j$ steps backward. Thus the overall probability for a sequence of steps occurring is $f^j b^{N-j}$. The sum of all the possible combination of random steps is $(b + f)^N$. Because $b + f = 1$, this sum is equal to 1, but we like to know how the components of the sum depend on *b*, *f* and N. The binomial expansion of the sum yields

$$(b + f)^N = b^N + Nb^{N-1}f + \{N(N-1)/2!\}b^{N-2}f^2 + \ldots + \{N!/j!(N-j)!\}b^{N-j}f^j + \ldots + f^N$$

The coefficients of each term in the expansion gives the statistical weight of the corresponding outcome. For example, the coefficient, $N!/\{j!(N-j)!\}$ represents the number of different combination of steps with the number of forward steps *j*, i..:

$$W_j = N!/\{j!(N-j)!\}$$

This is equivalent to the degeneracy of each state in the noncooperative structural transition for the two-state model. The probability of having *j* forward steps and therefore $N - j$ backward steps, P_j is

$$P_j = [N!/\{j!(N-j)!\}]f^j b^{N-j}$$

Similarly, this expression is equivalent to the probability of observing any state j in the two-state noncooperative transition. Knowing the probability of a random walk of N steps with j steps forward, the average number (mean value), $<j>$ and the mean of the square of the number, $<j^2>$ of steps forward can be calculated:

$$\langle j \rangle = \sum_{j=0}^{N} jP_j \qquad \text{and} \qquad \langle j^2 \rangle = \sum_{j=0}^{N} j^2 P_j \qquad \text{respectively.}$$

Conceptually and mathematically, the one-dimensional random walk is equivalent to the two-state noncooperative transition, with a forward step being one state and the backward step being the second state for each residue.

A quantity frequently used to express the average dimension of a flexible biomacromolecular structure is the root-mean-square (rms) end-to-end distance $<L^2>^{1/2}$ or alternatively rms radius $<R^2>^{1/2}$. These quantities are related to the rms displacement $<d^2>^{1/2}$. For an ensemble of completely random walks, the mean displacement, $<d>$, which measures the average net displacement after N steps of a random walk, would be zero. However, the mean-square (ms) displacement, $<d^2>$ and therefore the rms displacement $<d^2>^{1/2}$, which is a measure of the average absolute displacement from the origin at the end of the random walk, would not be zero. For example, a random walk ending five steps in front of ($d = +5$) and five steps behind ($d = -5$) the origin, the average net displacement is $<d> = 0$ and the rms displacement is $<d^2> = \{(+5)^2 + (-5)^2\}/2 = 25$. The rms displacement of $<d^2>^{1/2} = 5$ indicates only that this event of the random walk all end five steps from the origin. If the pace of each step is l and after N steps, the ms displacement and rms displacement respectively are

$$\langle d^2 \rangle = \sum_{j=0}^{N} l_j^2 = Nl^2 \qquad \text{and} \qquad \langle d^2 \rangle^{1/2} = (\Sigma l_j^2)^{1/2} = N^{1/2} l$$

Therefore rms displacement is dependent on $N^{1/2}$, i.e. the square root of the number of residues in a biomacromolecular chain. These relationships derived in a one-dimensional random walk can be extended to two and three dimensions, and thus are applicable to the description of the properties of the tertiary structures of biomacromolecules. The rms displacement describes the average absolute distance between the first and the last residues of a macromolecular chain. This is defined as the rms end-to-end distance $<L^2>^{1/2}$ that reflects the flexibility of the chain:

$$<L^2>^{1/2} = <d^2>^{1/2} = Nl^{1/2}$$

where N and l are number of residues and the distance between residues (normally the bond length constrained by the bond angle). For biomacromolecules, N is the number of functional units or segments of a chain that appears to behave randomly and l is the effective length of these random segments. The length of each of these segments defines the persistence length.

A quantity closely related to rms end-to-end displacement but a better measure of the effective size of a random yet compactly folded macromolecule is the rms radius $<R^2>^{1/2}$. This reflects the average displacement of each residue from the center of mass of a macromolecule and is a function of the mass of each residue, m_j and the distance of that residue from the center r_j according to:

$$<R^2>^{1/2} = [(\Sigma m_j r_j^2)/\Sigma m_j]^{1/2}$$

Assuming all masses are approximately equal and analogous to the rms end-to-end displacement, it can be shown that

$$<R^2>^{1/2} = (Nl^2/6)^{1/2}$$

for an open-ended (linear) random coil, and

$$<R^2>^{1/2} = (Nl^2/12)^{1/2}$$

for a circular random coil. Thus a biomacromolecule with $<R^2>^{1/2}$ proportional to $N^{1/2}$ or the square root of the molecular weight is said to behave as a random coil or a Gaussian chain.

9.4 STRUCTURAL TRANSITION: EXAMPLES

9.4.1 Coil-helix transition in polypeptides

The structural transition between a random coil and a helix is an important process in the folding pathway for proteins. Random coil refers to polymeric chains that are unfolded (or denatured) relative to the regular secondary structures and the compact tertiary structures (or native structures) of biomacromolecules. Polypeptide chains can adopt the α-helical conformation in which the carbonyl oxygen of residue i is hydrogen bonded to the amide hydrogen of residue i + 4 with optimal ϕ, ψ rotational angles of −50° (right-handed helix) and +50° (left-handed helix). According to the two-state model, the symbol, c (*a*, coil) represents a nonhydrogen bonded carbonyl oxygen and the symbol *h* (*b*, helix) if it is hydrogen bonded. *In vitro* coil-helix transition of polypeptides can be induced by variations in solution parameters such as temperature, pH and solvent composition. In a temperature-dependent coil-helix transition experiment, which shows a sharp inflection with the midpoint,

$$\underset{s_1}{\cdots cccccccc\cdots} \overset{s}{\longleftrightarrow} \underset{s_2}{\cdots cccchcccc\cdots}$$

the temperatures-dependent variable, s is treated as an equilibrium constant according to

$$\ln s = \ln(s_2/s_1) = -\Delta H^\circ/RT + c$$

The midpoint temperature of the transition, T_m is the temperature at which $s_1 = s_2$, i.e. $\ln s = \text{in}\ (1) = 0$ and $c = \Delta H^\circ/RT_m$. The temperature dependence of s can be expressed as

$$\ln s = (-\Delta H^\circ/R)(1/T - 1/T_m)$$

Thus the temperature-dependent transition of polypeptide chains can be modeled by the statistical mechanics treatment of the two-state zipper model.

In the structural transition experiments, which are commonly carried out in the denaturation of native proteins (reverse transition from helix to coil), an identical treatment is applicable. A polypeptide with N residues in a helical form undergoes helix-coil transition:

$$\begin{tikzcd}
 & \begin{matrix} chhhhhhhh\cdots \\ (w_{N-1} = \sigma s^{N-1}) \end{matrix} \quad \text{or} \quad \begin{matrix} \cdots hhhhhhhhc \\ (w_{N-1} = \sigma s^{N-1}) \end{matrix} \quad \text{initiation at terminus} \\
\begin{matrix} \cdots hhhhhhhh\cdots \\ (w_N = \sigma s^N) \end{matrix} \arrow[ur] \arrow[dr] & \\
 & \begin{matrix} \cdots hhhchhhh\cdots \\ (w_{N-1} = (N-2)\sigma^2 s^{N-1}) \end{matrix} \quad \text{initiation at the middle}
\end{tikzcd}$$

If either of the terminal residue initiates the conversion from *h* to c, the degeneracy of this state is 2 (initiation from either N- or C-terminus) and the statistical weight for the new state is $w_{N-1} = 2\sigma s^{N-1}$. However, if the transition initiates in the middle of the chain at residue j, the polypeptide is separated into two helical regions of j − 1 and N − j in length. Each separated helical segment has its own statistical weight; σs^{j-1} for the segment at N-terminus and σs^{N-j} for the C-terminus segment. The overall statistical weight for this state is $w_{N-1} = (N - 2)\sigma^2 s^{N-1}$, where N − 2 expresses the possible ways of placing a single residue as *h* in the chain without the two terminal residues. The σ^2 factor accounts for the need of the two continuous stretch of helix to have its own nucleation parameter. Thus the probability of melting a single residue at one end of the helix P_e versus that at a middle residue P_m in the helix-coil transition (denaturation) of proteins, is given by

$$P_e/P_m = \{2\sigma s^{N-1}\}/\{(N - 2)\sigma^2 s^{N-1}\sigma\} = 2/\{(N - 2)\sigma\}$$

The structural transition (coil-helix/helix-coil transition) of polypeptides shows the following features:

- Both the coil-to-helix and helix-to-coil transitions are cooperative and can be described by the two-state zipper model.
- In applying the zipper model, the transitions are characterized by the nucleation parameter σ and propagation parameter s.
- The parameters, σ and s are dependent on the nature of transition and the sequences of polypeptides. This is the basis for predicting secondary structures from the amino acid sequences of proteins.
- The cooperative transition is the result of a small σ ($\sigma \ll 1$), because initiation of an helix (or coil) is more difficult than its subsequent propagation.
- The cooperativity of the transition increases with increasing number of residues N of the polypeptide chains. The nucleation term dominates the transition in short polypeptides but the propagation parameter becomes dominant toward the end of the transition as N increases.
- For short polypeptides, the transition initiates at the ends and migrates toward the middle of the chain, but the helical or coil regions are expected to be present in the middle of the molecules for long polypeptides because there are so many nucleation sites in the middle of the chains.
- A sharp temperature dependent transition is observed for polypeptides with large N. The polypeptide changes from 80% in the coil form to 80% in the helix form (or vice versa) in a temperature range of ~7°. The transition is broader for the shorter polypeptides.

9.4.2 Helical transition in nucleic acids

Double-stranded DNA is held together by the interactions between complementary bases in two single strands of DNA, thus the helix-coil transition in a double-stranded nucleic acid is similar to that in a polypeptide with several new features in nucleic acids:

- *The dependence of σ on j*: The farther the separation (j, distance between the two helical segments) of the two base pair (bp) segments, the more difficult it is to bring the two together to form contiguous base pairs. Therefore, unlike the case for polypeptides, σ for nucleic acids is not a constant for a function of j. The first equilibrium constant is $\sigma_1 s$ and that for the subsequent one is $\sigma_j s$.

- *The participation of two complementary chains*: The formation of the very first base pair involve moving two complementary chains which are free to move independently in solution together. Therefore the equilibrium constant is taken as κs with κ expected to be much less than 1 and to be dependent on total concentration. The formation of subsequent bps from the two chains is constrained by the initial pairing.
- *Specific base pairings*: There are two major types of bps in nucleic acids; A·T(U) and G·C pairs. Because the stabilities of the two kinds of bps are different, generally the parameters s_A and s_G for A·T(U) and G·C pairs respectively, rather than a single s. The s_A and s_G are further dependent on at least the nearest-neighbor base sequence.

An inclusion of the above considerations into the analysis of the helix-coil transition of nucleic acids yields the following:

- When N is large, the transition from the completely helical structure to the completely coiled structure occurs within a narrow temperature range. This cooperative process is referred to as melting of a nucleic acid with characteristic melting temperature, T_m that corresponds to the midpoint of the transition.
- The enthalpy change for the reaction, . . . *cccccccc* . . . $\leftarrow$ s $\rightarrow$ . . . *ccchcccc* . . . is negative. In most of solvents, $s_G > s_A$; thus T_m for a nucleic acid rich in G + C is higher than that for a nucleic acid rich in A + T(U). If the base composition of a nucleic acid is intramolecularly heterogeneous, i.e. some segments are richer in A + T(U) than other segments, the melting profile is broadened, with the melting of the region richer in A + T(U) preceding the melting of the regions rich in G + C.
- For a nucleic acid of large molecular weight, T_m is expected to be independent of concentration. For short helices, T_m is lower at lower concentration. Similarly at the same total concentration of nucleotides, T_m for a high-molecular-weight nucleic acid is higher than that of a low-molecular-weight nucleic acid of the same base composition.
- If N is small, the coil regions should be at the end of the molecule. This is referred to as 'melting from the ends.'
- The processes of annealing (renaturing) and melting (denaturing) of DNA or RNA duplexes show hysterisis. The midpoint temperature for melting T_m is higher than the annealing temperature T_a. The melting starts as a well-defined structure and proceeds along a specific pathway, whereas the annealing must sample many possible thermodynamically degenerate forms at each state. This greatly increases the entropy of each annealed state thus a half-renatured duplex is much more heterogeneous (broader temperature dependent) in composition. Consequently the T_m and the cooperativity for melting is always greater than the T_a for renaturation.

The helical transitions from B-DNA to A-DNA and from B-DNA to Z-DNA can be treated as the zipper model. In these transitions, σ defines the extra energy ΔG_j^o required to form a junction between the two forms. A relatively small ΔG_j^o for the nucleation is probably due to the small difference in the bp stacking between A-DNA and B-DNA. The parameter s is associated with the difference in energy between bps in A-DNA versus B-DNA and on environment factors such as temperature, salt and organic solvents, all of which stabilize left-handed Z-DNA.

9.4.3 Topological transition of closed circular DNA duplex

In eukaryotes, the DNA is negatively coiled around histones to form nucleosomes. The prokaryotic nucleoid and plasmid DNA are also in a negatively supercoiled state. In addition, the positive (ahead) and negative (behind) supercoils are induced during RNA polymerase catalyzed transcription (Lui and Wang, 1987). The transitions between various helical conformations in double-stranded closed circular DNA (ccDNA) are dependent on the topological state of the DNA. The effect of supercoiling on a DNA transition is determined by studying the behavior of the topoisomers of ccDNA at discrete values of L ($L = T + W$ where L, T and W are linking number, twist and writhe respectively), which defines the topology of ccDNA. The helical transitions in ccDNA can be modeled by statistical thermodynamics by accurately describing the energies of the various topological states. The reference state is a relaxed (without supercoils) closed circular B-DNA with $W_0 = 0$ and $T = N$.

Any transition to a new state is accompanied by a change in twist $\Delta T = T_j - T_o$ and a concomitant change in writhe $\Delta W = W_j - W_o = -\Delta T$. Each bp j that is converted defines a new state of the system with a statistical weight w_j, which is dependent on the free energy $\Delta G^o(\Delta L)$ of the system (ΔL represents the topological state of ccDNA topoisomer relative to the reference state). The free energy of any topoisomer with a defined ΔL can partition between the unwinding of the DNA helix, ΔT and supercoiling, ΔW as given by

$$\Delta G^o(\Delta L) = \Delta G^o(\Delta T) + \Delta G^o(\Delta W) = \Delta G^o(\Delta T) + K\Delta W^2$$

$\Delta G^o(\Delta W)$ is the free energy of the superhelicity of the ccDNA of N bps and is defined as $\Delta G^o(\Delta W) = K\Delta W^2$ with the proportionality (spring) constant, $K = 1000\,RT/N$ for a ccDNA of N bps. This corresponds to one supercoil in a ccDNA of 1100 bp in length having a supercoiling free energy of 4 kJ/mol. The dependence of $\Delta G^o(\Delta W)$ on ΔW^2 implies that the lowest energy state has $\Delta W = 0$, and the introduction of any supercoiling, whether positive or negative, results in a higher energy state.

9.5 STRUCTURE PREDICTION FROM SEQUENCE BY STATISTICAL METHODS

9.5.1 Approaches

An acceptance of the basic principle that the amino acid sequence contains sufficient information to define the 3D structure of a protein in a particular environment has led to many diversified efforts to predict the conformation of proteins from amino acid sequences. Methods for abstracting biomacromolecular structures, in particular protein structures (section 16.5), from their sequences can be aimed at the secondary structure, initially following three stages:

1. prediction of secondary structure(s) from amino acid sequence;
2. determine approximate tertiary fold by packing the secondary structure(s) and
3. refinement of the tertiary fold to yield native conformation.

There are four approaches to secondary structure prediction (Fasman, 1989):

1. Empirical statistical methods that use parameters derived from known 3D structures (Chou and Fasman, 1974; Garnier *et al.*, 1978).
2. Methods based on correlation between physicochemical properties of monomeric residues and structures (Grantham, 1974; Deleage and Roux, 1987).

3. Methods based on prediction algorithms that use known structures of homologous sequences to assign secondary structures and folds (Levin *et al.*, 1986; Hilbert *et al.*, 1993).
4. Molecular mechanical methods that use FF parameters to model molecular structures (Kuntz *et al.*, 1976; Nemethy and Scheraga, 1977; Pittsyn and Finkelstein, 1983).

The theoretical approach, based on the assumption that a biomacromolecule folds so as to minimize the free energy of the system, leads to the development of potential functions that define the force field (FF). Chain folding is simulated computationally directed by surface gradients to find the energy minimum in EMin. Molecular dynamics is performed by integrating the equation of motion over time and conformational search is carried out by varying dihedral angles over the conformation space. These techniques were discussed in the previous sections on molecular modeling.

The heuristic approach (Stagle, 1971), based on the wealth of structural databases, proves to be most promising. In the pattern-based approach, the sequence alignment is performed to identify homologous structures from the database. Homology modeling or threading based on the known 3D structures of the homologues is then performed. The algorithms and executions of these analyses are now accessible on Web sites and will be discussed in the Chapter 16.

A number of physical-chemical properties such as molecular volume, exposure or accessible surface, hydrophobicity/hydrophilicity (polarity), charge or pK, hydrogen bonding potential, etc has been made to correlate with their relatedness between sequences and structures. Some of these properties for amino acids are accessible from AAindex database (http://www.genome.ad.jp/dbget). There are some overlaps between the physical-chemical properties and conformation preferences for amino acid residues derived from the known 3D structures of proteins. For example, on average, one third of the charged residues in a protein participate in the ion pair formation (ion pair is defined as the interaction between oppositely charged groups within the distance ≤4 Å) and more than $^3/_4$ of these are concerned with stabilizing the tertiary structure with less than 20% of ion pairs being buried. The hydrophobicity/hydrophilicity and accessibility of amino acids have been correlated with the packing of residues of a protein molecule since there are significant changes in the conformational preferences of the residues in going from the interior to the exterior of proteins.

The knowledge based structural prediction (Blundell *et al.*, 1987) depends on analogies between a biomacromolecule of known sequence and other biomacromolecules of the same class with known 3D structure at all levels in the hierarchy of biomacromolecular organization. In the numerically based statistical methods, the structural rules and parameters (conformational propensities) for each residue are extracted by statistical analyses of the structural database and used to predict the structure along the sequence of the macromolecule. Examples of the statistical methods that are commonly applied to predict secondary structure and folding preference of proteins will be illustrated.

9.5.2 Secondary structure of proteins and beyond

The requirements for hydrogen bond preservation in the folded structure result in the cooperative formation of hydrogen-bonded secondary structure regions in proteins. The secondary structure specifies regular polypeptide chain folding patterns of helices, sheets, coils and turns that are combined/folded into tertiary structure. Studies of two-state structural transition suggest that a statistical method can be developed to predict the probability (propensity) for sequences along polypeptide chains, to adopt specific secondary structures once the nucleation parameter (σ) and propagation parameter (s) are determined for all amino acids for all possible secondary structures. These parameters are environment-dependent, but attempts have been made to assign the average propensities for amino acids to adopt specific secondary structures in proteins.

In the statistical method of Chou and Fasman (Chou and Fasman, 1978), the propensities for a residue type to adopt three structural states; α-helix ($<P_\alpha>$), β-sheet ($<P_\beta>$) and turn ($<P_t>$) conformations are calculated for all of the 20 amino acids according to

$$<P_\alpha> = X_{\alpha i}/<X_\alpha>, \qquad <P_\beta> = X_{\beta i}/<X_\beta>, \qquad \text{and} \qquad <P_t> = X_{ti}/<X_t>$$

where $X_{\alpha i}$, $X_{\beta i}$ and X_{ti} are numbers of amino acid type *i* in respective α-helix, β-strand, and turn among the total number of amino acid type *i* in the data set analyzed. $<X_\alpha>$, $<X_\beta>$ and $<X_t>$ are average values of $X_{\alpha i}$, $X_{\beta i}$ and X_{ti} for all 20 amino acids respectively. Following steps are then undertaken to assign secondary structures along a peptide chain:

1. Assign helix and sheet propensity values and symbols (H, h, I, i, B, b for strong versus weak formers, indifferences and breakers).
2. Search for nucleation sites of α-helix and β-sheet:
   a. Helix: A 6-residue peptide containing at least four helix formers (H_α or h_α), where I_α counts as $^1/_2 h_\alpha$ and not more than one helix breaker (B_α or b_α).
   b. Sheet: A 5-residue peptide with at least three sheet formers (H_β or h_β) and not more than one sheet breaker (B_β or b_β)
3. Resolution of simultaneous helix-sheet assignments: For the same residues predicted to be α and β nucleation sites, calculate the average $<P_\alpha>$ and $<P_\beta>$ for these residues. The nucleation site is assigned to α or β with higher probability.
4. Propagation: Extend helix and sheet from the nucleation sites in both directions until the average probability of the end tetrapeptide falls below 1.0. End residues, which are breakers, are not to be included in the secondary structure.
5. Prediction of β turns: The frequencies for each amino acid type to be in the four-residue β turns (f_i, f_{i+1}, f_{i+2} and f_{i+4}) are calculated from the data set. For each peptide not assigned to α-helix and β-sheet, calculate the turn probability (π) as the product of four contiguous residues,

$$\pi = \prod_{i=1}^{4} f_i. \text{ The turn is predicted if } \pi > 4.6 \times 10^{-5}.$$

Another familiar statistical method using the information approach collects information on the significant pair-wise dependence of an amino acid in a given position with number of residues defined by the window (e.g. 7–12 residues) on either side of it in a particular secondary structural setting. The method of Garnier, Osguthorpe and Robson (GOR) (Garnier *et al.*, 1978) considers the effect that residues of a given position within the region eight residues N-terminal to eight residues C-terminal have on the structure of that position. Thus a profile that quantifies the contribution of the residue type toward the probability of one of four states; α-helix (H), β-sheet (E), turn (T) and coil (C), exists for each residue type. Four probabilities are calculated for each residue in the sequence by summing information from the 17 local residues, i ± 8. The statistical information derived from proteins of known structure is stored in four (17 × 20) matrices for H, E, T and C.

For any residue, the predicted state is the one with the largest probability value with reference to the known structures. The refinement of this approach includes the information from the alignment of homologous sequences (Garnier *et al.*, 1996).

The statistical methods for predicting secondary structures of proteins from amino acid sequences are widely practiced among investigators in biochemistry and can be accessed at Network Protein Sequence Analysis (NPS@) via http://npsa-pbil.ibcp.fr

The following characteristics of secondary structures may correlate to the sequence preference and subsequent fold of globular proteins:

- A great majority of α-helices is at the surface of proteins, approximately half exposed and half buried.
- Parallel β structures are typically buried on both sides, so that they appear hydrophobic near the center of a strand and hydrophilic near each end of the strand.
- Antiparallel β structures usually have one side buried and one side exposed, so there is a hydrophobic-hydrophilic periodicity with a repeat of about two residues for each strand.
- Turns are almost always exposed at the surface, therefore they have a tendency to occur at local maxima of hydrophilicity.
- Nonrepetitive loops are also exposed on the surface and generally hydrophilic.
- Completely disordered structures are the most hydrophilic of all conformations.
- Tails at N- and C-termini of a protein are likely to be disordered if they do not include large aliphatic or aromatic hydrophobics and a large proportion of charges, hydrogen bonding hydrophilics.
- An average residue buries about half of its potential accessible surface ($\sim 80\,\text{Å}^2$) in going from an extended conformation to an α-helix or two stranded β-sheet. On folding, both polar and nonpolar surfaces are reduced by a similar amount, approximately three-quarters.

The hydrophobicity versus hydrophilicity of side chains are important for predicting conformations of proteins by virtue of their preferential occurrence at the interior versus exterior of protein molecules. Virtually all ionized residues are on the surface of the molecules exposed to the solvent. Those hydrophilic residues, which are in the protein interior, are either ion-paired or hydrogen-bonded. In general, hydrophobic residues tend to be buried inside whilst hydrophilic residues prefer to be at the surface of protein molecules. The spontaneous folding of a polypeptide chain into a compact globular protein depends to some extent on association of water or avoidance of it by amino acid residues of the chain, therefore it is expected that the sequence of amino acid hydrophobicity may influence the folding of the protein molecule. The numerical approach to measure and assign the hydrophobicity of amino acid residues is difficult and complicated. No unified scale has been emerged though attempts have been made in this direction. Table 9.3 lists some representative hydrophobicity scales/indices.

The computational approaches to predict protein structures from their amino acid sequences can be either empirical or knowledge-based. The former involves calculations of the energetically favorable structures by the use of the parameterized force fields (section 9.2) and the latter performs statistical analysis with references to the existing structures employing ranges of bioinformatic techniques (section 16.5). The comparative structural analysis leads to homologous models, which are then refined by energy minimization.

TABLE 9.3 Hydrophobicities and related parameters of amino acid residues

Residue	Transfer energy[a] (kJ/mol)	Hydration.[b] (kJ/mol)	Partition coeff.[c] (kJ/mol)	Hydropathy index[d]	Memb.-buried prefer.[e]	Consen. scale[f]	Access. Surface[g] (Å^2)	Buried residues % mole f[h]
Gly	0.00	9.9	0.00	–0.4	0.62	0.16	75	11.8
Ala	–2.09	8.0	1.76	1.8	1.56	0.25	115	11.2
Val	–6.28	8.20	6.84	4.2	1.14	0.54	155	12.9
Leu	–7.53	9.39	9.56	3.8	2.93	0.53	170	11.7
Ile	–7.53	8.86	10.1	4.5	1.67	0.73	175	8.6
Pro	–5.85		4.01	–1.6	0.76	–0.07	145	2.7
Cys	–4.18	–5.11	5.52	2.5	1.23	0.04	135	4.1
Met	–5.44	–6.1	6.92	1.9	2.96	0.26	185	1.9
Thr	–1.67	–20.1	1.44	–0.7	0.91	–0.18	140	4.9
Ser	1.25	–20.9	–0.21	–0.8	0.81	–0.26	115	8.0
Phe	–10.5	–3.1	10.1	2.8	2.03	0.61	210	5.1
Trp	–14.2	–24.2	12.7	–0.9	1.08	0.37	255	2.2
Tyr	–9.62	–25.2	5.40	–1.3	0.68	0.02	230	2.6
Asn	0.836	–39.9	–3.43	–3.5	0.27	–0.64	160	2.9
Gln	0.836	–38.7	–1.20	–3.5	0.51	–0.69	180	1.6
Asp	10.5	–45.1	–4.33	–3.5	0.14	–0.72	150	2.9
Glu	10.5	–42.2	–3.64	–3.5	0.23	–0.62	190	1.8
His	–2.09	–42.3	0.75	–3.2	0.29	–0.40	195	2.0
Lys	12.5	–39.2	–5.56	–3.9	0.15	–1.1	200	0.5
Arg	12.5	–82.1	–5.64	–4.5	0.45	–1.8	225	0.5

Notes: Data taken from [a]Nozaki and Tanford (1971) and Levitt (1976), calculated from solubilities in water and ethanol or dioxane; [b]Radzicka and Wolfenden (1988), transfer free energies from gas phase to water; [c]Fauchere and Plisk (1983), calculated from partition from octanol to water; [d]Kyte and Doolittle (1982); [e]Argos *et al*. (1982), from 1125 amino acids found in protein segments within membranes; [f]Eisenberg *et al*. (1982); [g]Chothia (1976), for residue R in the extended tripeptide Gly-R-Gly; [h]Janin (1979), buried residues are those residues with an accessible surface area of less than 20 Å^2.

9.5.3 Functional sites of proteins

Enzyme active sites commonly occur in large and deep clefts on the protein surface and the need for significant favorable interactions between ligand and protein usually means that ligands also bind in surface depression. In analogy to pharmacophore modeling in computer-aided drug design (Tropsha and Zheng, 2001), two major strategies to predict functional sites are commonly followed:

1. *Protein-based approach*: Proteins sharing a high degree of sequence homology show a high degree of similarity in 3D structures. Furthermore, important functional sites in proteins usually display a high level of conservation. Therefore the prediction of functional sites using sequence/structure similarity is a very powerful tool. The functional site sequences can be readily obtained by statistical analysis/sequence alignment of homologous proteins with known identical functions. This is the knowledge-based approach and its facilities are available at various Web servers (subsection 16.2.3). Approaches have also been developed to identify functionally important regions on protein surfaces (Landgraf *et al.*, 2001; Lichtarge and Sowa, 2002).
2. *Ligand-based approach*: At least two reasons for an alternative strategy to predicting functional site can be contemplated. First, proteins with different folds may share

aspects of a function that is reflected in site similarity independent of evolutionary homology. Second, within evolutionary families, there are often functional differences among proteins of the same structural fold, based on their sequences and structures (Nagano *et al.*, 2002). Therefore mapping of functional site residues to structural alignments of protein family members may be very useful. There are two distinguishable approaches to functional site similarity (Campbell *et al.*, 2003):

1. The template approach involves the creation of 3D templates reflecting particular functions or defining the recognition properties of ligand binding sites. Such templates may focus on residues/chemical groups flanking the binding pocket or binding site volume.
2. Similarity search conducts systematic search of functional site databases for chemical groups able to make ion pairs/hydrogen bonds and/or hydrophobic interactions with ligands from which representative sites as surfaces with electrostatic and hydrophobic characteristics are generated (Schmitt *et al.*, 2002).

It is important to recognize similarities in functional sites. However, differences are also important as they are often related to specificity. Many protein families share details of molecular function but vary in finer details such as their substrate specificity. Therefore those residues that are involved in discerning subclasses are likewise important in defining functional sites (Hannenhalli and Russell, 2000). For examples, the nicotinamide nucleotide binding core, $(\alpha_2\beta_3)_2$ are common to dehydrogenases but the key residue, Asp/Glu for NAD^+ is replaced by Arg for $NADP^+$ dependent enzymes. Similarly, key catalytic residues are common to all protein kinases but two regions of the sequence differ and are known to confer the specificity for either Ser/Thr or Tyr specific enzymes.

9.5.4 Nucleic acid fold

The nucleotide sequence of a nucleic acid immediately provides plausible models for the approximate secondary structures. The base-pairing patterns, A-T(U) and G-C, are easy to identify and several simple rules can guide selection of folding patterns:

- Long continuous stretches of complementary bps will almost always form double-stranded hairpins.
- One long duplex region is preferable to two or more shorter separate regions with a total duplex length equal to the long region.
- For a choice between two equivalent and mutually exclusive pairings, the one with the highest G-C content is the more stable. If more base pairings have equal G-C content, the one between residues closer in the primary structure is more stable.

The bp complementarity in nucleic acids is unique among biomacromolecules. The effectiveness of these rules highlights a major difference between nucleic acids versus proteins and glycans. There are no simple patterns of complementarity that allow secondary structures of proteins to be assigned merely by considering pairwise interactions between residues, whilst every residues in glycans are potentially capable of participating in the hydrogen bonded secondary structures.

In RNA, the original set consisting of the Watson–Crick base pairs is complemented by the G-U wobble pair, which is admissible in RNA double helices. Other admissible bps include U-U in internal loops as well as A-A, G-A or G-G (purine-purine closing pairs) at the ends of double helical regions or in multiloops. The secondary structures, which can be drawn in two dimensions without knots or pseudoknots, are indispensable for the predictions of 2D structures of RNA (Schuster *et al.*, 1997). Two general approaches are available for the prediction of RNA structures in particular base pairings.

1. The thermodynamic approach computes for minimum free energy structures by dynamic programming, including suboptimal structures (Zuker, 1989), generic algorithms (Gultyaev *et al.*, 1995) and simulated annealing (Schmitz and Steger, 1995).
2. The evolutionary approach performs comparative sequence analysis leading to models of the phylogenetic/physiologically active structures (Gutell, 1993).

An incorporation of noncanonical bp, such as G-U wobble pairs, U-U in internal loops, purine-purine bps at the ends of double helical stacks, base triplets, knots and pseudoknots approximates tertiary interactions. One approach to molecular modeling of RNA structures, known as macromolecular conformations by symbolic programming (MC-SYM), is based on the symbolic creation of coarse structures that are refined by EMin/MolD calculations. Using experimental constraints in the structural modeling greatly improve the computational outcomes. The prediction of 3D biomacromolecular structures based on the computation of minimal potential energies faces a formidable task because of the enormously large numbers of local optima. A method based on conformational searches using a genetic algorithm (Koza, 1993), followed by refinement via Emin, is an attractive approach.

The task of assigning a plausible pattern of base pairing is greatly simplified if sequences are available for different species of RNA known to possess similar structures and functions. For example, the cloverleaf structure and the L-shaped fold have served as good approximations for modeling the secondary and tertiary structures of tRNA respectively. DNA and RNA sequences can be submitted to respective DNA mfold (http://bioinfo.math.rpi.edu/~mfold/dna) and RNA mfold (http://bioinfo.math.rpi.edu/~mfold/rna) for fold predictions.

9.6 MOLECULAR DOCKING: PREDICTION OF BIOMACROMOLECULAR BINDING

One basic property of biomacromolecules is their ability to interact specifically with small molecule ligands and biomacromolecules. The binding of ligands to receptor biomacromolecules is central to numerous biological processes. Therefore the prediction of the binding modes between the ligand receptor protein (docking problem) is of fundamental importance to computational biochemistry. Assuming the protein structure is available from Protein Data Bank (PDB) at http://www.rcsb.org/pdb/, the primary challenge is to predict ligand orientation (molecular docking) and binding affinity. Ligand-receptor interaction is an important initial step in protein function. The structure of ligand-receptor complex profoundly affects the specificity and efficiency of protein action. Molecular docking explores the binding modes of two interacting molecules, depending upon their topographic features or energy-based consideration and aims to fit them into conformations that lead to favorable interactions (Campbell *et al.*, 2003). Thus one molecule (e.g. small molecule ligand) is brought into the vicinity of another (e.g. receptor biomacromolecule) while calculating the interaction energies of the many mutual orientations of the two interacting molecules to form ligand-receptor complex. In the complex, ligand and receptor molecules are presumed to adopt the energetically most favorable docking structures. Thus the goal of molecular docking is to search for the structure and stability of the complex with the global minimum energy.

Docking protocols consist of two components; a search strategy and a scoring function. The search algorithm should generate an optimum number of configurations that include the experimentally determined binding mode. A common approach in modeling molecular flexibility is to consider only the conformational space of the ligand, assuming a rigid receptor throughout the docking protocol. The scoring function should be able to distinguish the experimental binding modes from all other modes explored through the searching algorithm and ranks each conformation that is both accurate and efficient. Scoring methods can range from MM such as AMBER or CHARMm through to an empirical free energy scoring function (Eldridge *et al.*, 1997) or knowledge based functions (Muegge and Martin, 1999).

There are two classes of strategies for docking a ligand to a receptor. The first class uses a whole ligand molecule as a starting point and employs a search algorithm to explore the energy profile of the ligand at the binding site, searching for optimal solutions for a specific scoring function. The search algorithms include molecular dynamics and simulated annealing, MC optimization, genetic algorithms and evolutionary programming as well as a geometric complementary match. Representative examples are AutoDock, DARWIN, DOCK, GOLD and FTDOCK. The second class starts by placing one or several fragments (substructures) of a ligand into a binding pocket, and then constructs the rest of the molecule in the site (*de novo* design method). Representative examples are ADAM, DOCK4.0, FlexX, LUDI and SPECTTOPE. These tools are listed in Table 9.4.

In an interactive docking, an initial knowledge of the binding site is normally required. The ligand is interactively placed on to the binding site. Geometric restraints such as distances, angles and dihedrals between bonded or nonbonded atoms, may facilitate the docking process. Two parameters, the equilibrium value of the internal coordinate and the force constant for the harmonic potential need to be specified. For interatomic distance, the equilibrium restraint may be the initial length of the bond if it is desired to have a particular bond length remaining constant during a simulation. If we wish to force a bond coordinate to a new value, the equilibrium internal coordinate is the new value. Normally, the force constants are 29.3 kJ/mol Å^2 for an interatomic distance (larger for nonbonded distances), 52.3 kJ/mol Å^2 for an angle, and 66.9 kcal/mol Å^2 for a dihedral angle. Such constraints also ensure the confinement of the ligand molecule at the proximity of the binding site of the receptor during energy minimization. We may freeze most of the receptor molecule while allowing the ligand and contact residues to move in the field of the frozen atoms. Thus only the selected atoms (e.g. ligand molecule and contact residues) move, while other (frozen) atoms influence the calculation.

In an automatic docking, for example, a ligand or its substructure is allowed to fit into potential binding cleft of the receptor of known crystal structure. Initially the surface complementarity between the ligand and the receptor is determined by searching for a geometrical fit using the molecular surface as starting images (spheres). From each surface points, a set of spheres that fill all pockets and grooves on the surface of the receptor is generated, and various criteria are introduced to reduce their number to one sphere per atom. Within this approximation of both ligand and receptor shapes by sets of spheres, the search is made to fit the set of ligand spheres within the set of receptor spheres. The matching algorithm collects all fits that are possible by comparing internal distances in both ligand and receptor and lists pairs of ligands and receptor spheres having all internal distances matching within a tolerance value. Ligand atom coordinates are calculated and the locations of the ligand atoms are optimized to improve the fit.

There are numerous cases of proteins for which structures have been determined in more than one state of ligation. In some cases, the structures undergo little change, except perhaps for specific and localized changes associated with particular functional residues,

TABLE 9.4 Some tools (programs) for molecular docking

Program	Feature	Ref
ADAM	This is a FB algorithm for flexible docking *via* alignment of a fragment based on HB motifs, followed by energy minimization with AMBER. This includes matching the HB dummy atoms of the receptor with all potential HB atoms in the ligand fragment by superposition and minimization. The conformations of the rest of the ligand are scanned for the lowest energy conformer.	1
AutoDock	An early AutoDock program docks flexible ligands into the binding pocket of a rigid receptor using Metropolis MC simulated annealing with a grid based evaluation of the AMBER energy. The version 3.0 (AutoDock 3.0) uses GA as a global optimizer combined with energy minimization as a local search method. In this implementation, the GA uses two point crossover and mutation operators with AMBER as the score function.	2, 3
DARWIN	This program combines a GA and a local gradient minimization with CHARMM as the scoring function for flexible docking of protein complexes. The populations are modified by standard mutation and crossover operators.	4
DOCK	An early DOCK program searches for geometrically allowed ligand-binding modes by using the ligand and receptor cavity as sets of spheres, matching the sphere sets, and orientating the ligand. Having generated multiple orientations, an intermolecular score is calculated based on the AMBER where receptor terms are calculated on a grid. The version 4.0 (DOCK 4.0) is FBM in that a ligand anchor fragment is selected and placed in the receptor, followed by rigid body simplex minimization. The conformations of the remaining parts of the ligand are searched by a limited backtrack method and minimized. The scoring function incorporates an intramolecular score for the ligand.	5, 6
DockVision	This program uses MC to dock rigid ligand and rigid receptor by generating a random ligand orientation first. The MC and MC simulated annealing are applied. Ligand orientations are subjected to cluster analysis based on a RMSD score.	7
FlexX	The FB algorithm, FlexX selects the base fragment for ligand containing the predominant interactions with the receptor based on the torsion angle DB (MIMUMBA). An alignment is carried out to optimize the number of favorable interactions. Three sites on the fragment are mapped onto three sites of the receptor and the superposition of ligand triplets onto the receptor is carried out. The ligand is then built in an incremental fashion and the placements are ranked from which the best solutions are subjected to a cluster analysis. The highest rank solution from each cluster is then used in the next iterations until the complete ligand is built.	8
FLOG	Flexible Ligand Oriented on a Grid selects ligands from a DB, complementary to a receptor of known 3D structure. Distances between favorable sites of interaction in the protein are calculated along with atomic distances of the ligand. The ligand is then superimposed onto the protein interaction sites, optimized and scored with a function that includes van der Waals, electrostatic, HB and ψI.	9
FTDOCK	This is a rigid docking protocol based on shape complementarity with a model for the electrostatic interactions. A grid is placed over the protein and ligand of which each gp is designated as open space or inside the molecule. The protein is further divided into gps on the surface or buried within the molecule. The correlation function which is to be optimized by rigid body translation/rotation of the ligand, is the product of a gp in the ligand with the corresponding gp of the protein summed over all points. A high correlation score denotes good surface complementarity between the molecules.	10
GOLD	The program uses a GA search strategy and includes ligand flexibility and rotational flexibility for selected receptor hydrogens. The ligand-receptor HB is matched with a least squares to maximize the number of intermolecular HB for each GA move. The scoring function is the sum of a 4–8 intermolecular dispersion potential of HB term and a 6–12 intramolecular potential for the internal energy of the ligand.	11

TABLE 9.4 ***continued***

Program	Feature	Ref
CM	Internal Coordiantes Mechanics performs flexible ligand-protein docking using MC minimization method. The algorithm makes one of initial random moves (rigid body ligand move or torsion moves) to be followed by local minimization of ECEPP function with a conjugated gradient minimizer.	12
LIGIN	LIGIN uses a surface complementarity approach to dock rigid ligands into a rigid receptor. The complementarity function which is a sum of atomic surface contacts weighted according to molecular interactions, is maximized. The ligand structures generated are then optimized for HB distances between the ligand and protein.	13
LUDI	This FBM fits ligands into the active site of a receptor by matching complementary polar and ψ groups. The program operates by (1) calculating ligand and protein interaction sites that are discrete positions in space which can either form HB or ψI. These sites are derived from non-bonded contact distributions *via* CCDC DB, or from a set of geometric rules, or from the output of GRID, (2) fitting molecular fragments onto the interaction sites by RMSD superposition algorithm which fits complementary ligand groups onto the receptor with user defined interaction distance (3) joining of fragments using a DB of bridge fragments and fitting algorithm.	14
MCDOCK	Multiple stage strategy is applied to dock a flexible ligand to a rigid receptor by (1) placing the ligand in the binding site, (2) reducing the overlap of ligand and protein atoms by random moves, (3) performing metropolis MC with simulated annealing, (4) applying CHARMM based scoring function, and (5) MC simulation.	15
Prodock	The program docks flexible ligands to a flexible binding site (representing the structure with internal coordinates) using MC minimization technique in combination with local gradient-based minimization. Two FF, AMBER and ECEPP are implemented.	16
SANDOCK	This tool uses both shape and chemical complementarity to dock the ligand into the accessible surface of the protein binding site by a distance matching algorithm. The tolerance for the distance complementarity is adjusted to favor HB and ψI between the ligand and receptor.	17
SLIDE	In FB SLIDE, Anchor fragments are chosen which contain triplets of mps of which one is a HB donor/acceptor or ψ ring center. The best matches between triplets of mps on the ligand with mps on the protein are calculated. A least squares superposition algorithm is then used to place the fragment into the binding site. The rest of the ligand is added to the anchor fragment using the original DB conformation. Any intermolecular overlaps are removed and the complexes are scored based on HB and ψ complementarity.	18

Notes: 1. Abbreviations used: DB, database; FB(M), fragment-based (method); FF, force field; GA, genetic algorithms; gp, grid point; HB, hydrogen bond(s); ψ(I), hydrophobic (interactions); MC, Monte Carlo method; mp, matching point.

2. References cited are: [1] Mizutani *et al.* (1994), [2] Morris *et al.* (1996), [3] Morris *et al.* (1998), [4] Taylorand Burnett (2000), [5] Kuntz *et al.* (1982), [6] Ewing and Kuntz (1997), [7] Hart and Read (1992), [8] Rarey *et al.* (1996), [9] Miller *et al.* (1994), [10] Gabb *et al.* (1997), [11] Jones *et al.* (1997), [12] Abagyan *et al.* (1994), [13] Sobolev *et al.* (1996), [14] Böhm (1992), [15] Liu and Wang (1999), [16] Shoichet and Kuntz (1996), [17] Burkhand *et al.* (1998), [18] Schnecke *et al.* (1998), [GRID] Goodford (1983), [MIMUMBA] Klebe and Mietzner (1994), [RMSD superposition] Kabasch (1976).

e.g. triose phosphate isomerase. Other cases, such as citrate synthase, show conformational change in which binding of ligand in a site between two domains leads to a closure of the inter-domain cleft. In addition, the long-range integrated conformational changes associated with allosteric transitions of proteins must be considered. Thus protein structural flexibility is of fundamental importance to the applicability of docking methods. Docking with full protein flexibility is currently not feasible for a large number of ligands and therefore some level of approximation and constrains must be introduced (Hindle *et al.*, 2002; Schafferhans and Klebe, 2001; Zabell and Post, 2002).

9.7 REFERENCES

Abagyan, R., Totrov, M. and Kuznetsov, D. (1994) *Journal of Computer Chemistry*, **15**, 488–506.

Allen, M.P. and Tildesley, D.J. (1967) *Computer Simulation of Liquids*, Oxford University Press, New York.

Allinger, N.L., Yuh, Y.H. and Lii, J.H. (1989) *Journal of the American Chemistry Society*, **111**, 8551–66.

Altona, C. and Faber, D.H. (1974) *Topics in Current Chemistry*, **45**, 1–38.

Anfinsen, C.B., Haber, E., Sela, M. and White, F.H. (1961) *Proceedings of the National Academy of Sciences USA*, **47**, 1309–14.

Anfinsen, C.B. (1973) *Science*, **181**, 223–30.

Argos, P., Rao, J.K.M. and Hargrave, P.A. (1982) *European Journal of Biochemistry*, **128**, 567–75

Becker, O.M., MacKerell, A.D. Jr., Roux, B. and Watanabe, M. (eds) (2001) *Computational Biochemistry and Biophysics*, Marcel Dekker, New York.

Berkert, U. and Allinger, N.L. (1982) *Molecular Mechanics*, American Chemical Society, Washington, DC.

Blundell, T.L., Sibanda, B.L., Sternberg, M.J.E. and Thornton, J.M. (1987) *Nature*, **326**, 347–52.

Böhm, H.J. (1992) *Journal of Computer Aided Molecular Design*, **6**, 61–78.

Boyd, D.B. and Lipkowitz, K.B. (1982) *Journal of Chemistry Education*, **59**, 269–74.

Boyd, D.B. (1995) *Review of Computer Chemistry*, **6**, 383–417.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D. *et al.* (1983) *Journal of Computer Chemistry*, **4**, 187–217.

Brooks III, C.I., Karplus, M. and Pettitt, B.M. (1988) *Proteins: A Theoretical Prospective of Dynamics, Structure and Thermodynamics*, John Wiley & Sons, New York.

Burkhand, P., Taylor, P. and Walkinshaw, M.D. (1998) *Journal of Molecular Biology*, **277**, 449–66.

Campbell, S.J., Gold, N.D., Jackson, R.M. and Westhead, D.R. (2003) *Current Opinions in Structural Biology*, **13**, 389–95.

Chang, G., Guida, W.C. and Still, W.C. (1989) *Journal of the American Chemistry Society*, **111**, 4379–86.

Chou, P.Y. and Fasman, G.D. (1974) *Biochemistry*, **13**, 211–22.

Chou, P.Y. and Fasman, G.D. (1978) *Advances in Enzymology*, **47**, 45–148.

Clark, M., Cramer, R.D., III, and van Opensch, N. (1989) *Journal of Computer Chemistry*, **10**, 982-1-12.

Chothia, C. (1976) *Journal of Molecular Biology*, **105**, 1–14.

Deléage, G. and Roux, B. (1987) *Protein Engineering*, **1**, 289–94.

Eisenberg, D., Weiss, R.M., Terwilliger, T.C. and Wilcox, W. (1982) *Faraday Symposium Chemistry Society*, **17**, 109–20.

Eldridge, M.D., Murray, C.W., Auton, T.R. *et al.* (1997) *Journal of Computer Aided Modeling and Design*, **11**, 425–45.

Eriksson, L.A. (2001) *Theoretical Biochemistry*, Elsevier, New York.

Ewing, T.J.A. and Kuntz, I.D. (1997) *Journal of Computer Chemistry*, **18**, 1175–89.

Fasman, G.D. (ed.) (1989) *Prediction of Protein Structure and Principles of Protein Conformation*, Plenum Press, New York.

Fauchere, J.-L. and Plisk, V. (1983) *European Journal of Medical Chemistry*, **18**, 369–75.

French, A.D. and Brady, J.W. (eds) (1990) *Computational Modeling of Carbohydrate Molecules*, American Chemical Society, Washington, DC.

Gabb, H.A., Jackson, R.M. and Sternberg, M.J.E. (1997) *Journal of Molecular Biology*, **272**, 106–20.

Garnier, J., Osguthorpe, J.D. and Robson, B. (1978) *Journal of Molecular Biology*, **120**, 97–120.

Garnier, J., Gibrat, J.-F. and Robson, B. (1996) *Methods in Enzymology*, **266**, 540–53.

Gogonea, V., Uárez, D.S., van der Vaart, A. and Merz, K.M. Jr. (2001) *Current Opinions in Structual Biology*, **11**, 217–23.

Goodford, P.J. (1983) *Journal of Medical Chemistry*, **28**, 849–57.

Grantham, R. (1974) *Science*, **185**, 862–4.

Gultyaev, A.P., Van Batenburg, F.H.D. and Pleij, C.W.A. (1995) *Journal of Theoretical Biology*, **174**, 269–80.

Gundertofte, K. and Jørgensen, F.S. (eds) (2000) *Molecular Modeling and Prediction of Bioactivity*, Plenum Publishers, New York.

Gutell, R.R. (1993) *Current Opinions in Structual Biology*, **3**, 313–22.

Hart, T.N. and Read, R.J. (1992) *Proteins*, **13**, 206–22.

Hannenhalli, S.S. and Russell, R.B. (2000) *Journal of Molecular Biology*, **303**, 61–76.

Hehre, J.W., Radom, L., Schleyer, P. and Pople, J. (1986) Ab Initio *Molecular Orbital Theory*, John Wiley and Sons, Inc. New York.

Herman, J., Berendsen, H.J.C., van Gunsteren, W. and Postma, J.P.M. (1984) *Biopolymers*, **23**, 1513–8.

Hilbert, M., Bohm, G. and Jaenicke, R. (1993) *Proteins*, **17**, 138–51.

Hindle, S.A., Rarey, M., Buning, C. and Lengaue, T. (2002) *Journal of Computer Aided Modeling and Design*, **16**, 129–49.

Höltje, H.-D., Sippl, W., Rognan, D. and Folkers, G. (2003) *Molecular Modeling: Basic Principles and Applications*, VCH, Weinheim, Germany.

Howard, A.E. and Kollman, P.A. (1988) *Journal of Medical Chemistry*, **31**, 1669–75.

Jalaie, M. and Lipkowitz, K.B. (2000) *Reviews in Computer Chemistry*, **14**, 441–86.

Janin, J. (1979) *Nature*, **277**, 491–2.

Jones, G., Willett, P., Glen, R.C. *et al.* (1997) *Journal of Molecular Biology*, **267**, 727–48.

Jorgensen, W.L. and Tirado-Rives, J. (1988) *Journal of Medical Chemistry*, **31**, 1669–75.

Kabasch, W. (1976) *Acta Crystallography*, **A32**, 922–3.

Klebe, G. and Mietzner, T. (1994) *Journal of Computer Aided Molecular Design*, **8**, 583–606.

Koza, J. (1993) *Genetic Programming*, MIT Press, Cambridge, MA.

Kuntz, I.D., Crippen, G.M., Kollman, P.A. and Kimelman, D. (1976) *Journal of Molecular Biology*, **106**, 983–94.

Kuntz, I.D., Blaney, J.M., Oatley, S.J. *et al.* (1982) *Journal of Molecular Biology*, **161**, 269–88.

Kyte, J. and Doolittle, R.F. (1982) *Journal of Molecular Biology*, **157**, 105–32.

Landgraf, R., Xenarios, I. and Eisenberg, D. (2001) *Journal of Molecular Biology*, **307**, 487–502.

Leach, A.R. (1996) *Molecular Modeling Principles and Applications*, Longman, Essex, UK.

Leontis, N.B. and SantaLucia, J. Jr. (eds) (1998) *Molecular Modeling of Nucleic Acids*, American Chemical Society, Washington, DC.

Levin, J.M., Robson, B. and Garnier, J. (1986) *FEBS Letters*, **205**, 303–8.

Levitt, M. (1976) *Journal of Molecular Biology*, **104**, 59–107.

Lichtarge, O. and Sowa, M.E. (2002) *Current Opinions in Structual Biology*, **12**, 21–7.

Lui, L. and Wang, J.C. (1987) *Proceedings of the National Academy of Sciences USA*, **84**, 7024–7.

Liu, M. and Wang, S.J. (1999) *Journal of Computer Aided Molecular Design*, **13**, 435–51.

McCammon, J.A. and Harvey, S.C. (1987). *Dynamics of Proteins and Nucleic Acids*, Cambridge University Press, New York.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N. *et al.* (1953) *Journal of Chemical Physics*, **21**, 1087–92.

Miller, M.D., Kearsley, S.K., Underwood, D.J. and Sheridan, R.P. (1994) *Journal of Computer Aided Molecular Design*, **8**, 153–74.

Mizutani, M.Y., Tomioka, N. and Itai, A.J. (1994) *Journal of Molecular Biology*, **243**, 310–26.

Morris, G.M., Goodsell, D.S., Huey, R. and Olson, A.J. (1996) *Journal of Computer Aided Molecular Design*, **10**, 293–304.

Morris, G.M., Goodsell, D.S., Halliday, R.S. *et al.* (2000) *Proteins*, **41**, 173–91.

Muegge, I. and Martin, Y.C. (1999) *Journal of Medical Chemistry*, **42**, 791–804.

Nu, N., Orengo, C.A. and Thornton, J.M. (2002) *Journal of Molecular Biology*, **321**, 741–65.

Nemethy, G. and Scheraga, H.A. (1977) *Quarterly Reviews in Biophysics*, **3**, 239–352.

Nemathy, G., Porttle, M.S. and Scheraga, H. (1983) *Journal of Physical Chemistry*, **87**, 1882–7.

Nozaki, Y. and Tanford, C. (1971) *Journal of Biological Chemistry*, **246**, 2211–7.

Osawa, E. and Lipkowitz, K.B. (1995) *Reviews in Computer Chemistry*, **6**, 355–81.

Pittsyn, O.B. and Finkelstein, A. (1983) *Biopolymers*, **22**, 15–25.

Radzicka, A. and Wolfenden (1988) *Biochemistry*, **27**, 1664–70.

Rarey, M., Kramer, B., Lengauer, T. and Klebe, G. (1996) *Journal of Molecular Biology*, **261**, 470–89.

Schafferhans, A. and Klebe, G. (2001) *Journal of Molecular Biology*, **307**, 407–27.

Schlecht, M.F. (1998) *Molecular Modeling on the PC*, Wiley-VCH, New York.

Schmitt, S., Kuhn, D. and Klebe, G. (2002) *Journal of Molecular Biology*, **323**, 387–406.

Schmitz, M. and Steger, G. (1995) *Journal of Molecular Biology*, **255**, 254–66.

Shoichet, B.K. and Kuntz, I.D. (1996) *Chemical Biology*, **3**, 151–6.

Schnecke, V., Swanson, C.A., Getzoff, E.D. *et al.* (1998) *Proteins*, **33**, 74–87.

Schuster, P., Stadler, P.F. and Renner, A. (1997) *Current Opinions in Structural Biology*, **7**, 229–35.

Sharp, K.A., Nicolls, A., Friedman, R. and Honig, B. (1991) *Biochemistry*, **30**, 9686–97.

Sobolev, V., Wade, R.C., Vriend, G. and Edelman, M. (1996) *Proteins*, **25**, 120–9.

Stagle, J.R. (1971) *Artificial Intelligence: The Heuristic Programming Approach*, McGraw Hill, New York.

Sun, S.F. (1994) *Physical Chemistry of Macromolecules*, John Wiley & Sons, New York.

Tropsha, A. and Zheng, W. (2001) in *Computational Biochemistry and Biophysics* (edited by O.M. Becker, A.D. Mackerell, Jr, B. Rocex and M. Watanabe), Marcel Dekker, Inc., New York.

Tsai, C.S. (2002) *An Introduction to Computational Biochemistry*. John Wiley & Sons, New York.

Van Holde, K.E., Johnson, W.C. and Ho, P.S. (1998) *Principles of Physical Biochemistry*, Prentice Hall, Upper Saddle River, NJ.

Warshel, A. (1991) *Computer Modeling of Chemical Reactions in Enzymes and Solutions*, John Wiley, New York.

Webster, D.M. (ed.) (2000) *Protein Structure Prediction: Methods and protocols*. Humana Press, Totowa, NJ.

Weiner, S.J., Kollman, P.A., Case, D.A. *et al.* (1984). *Journal of the American Chemistry Society*, **106**, 765–84.

Weiner, S.J., Kollman, P.A., Nguyen, D.T. and Case, D.A. (1986) *Journal of Computer Chemistry*, **7**, 230–52.

Wiberg, K.B. (1965) *Journal of the American Chemistry Society*, **87**, 1070–8.

Zabell, A.P. and Post, C.B. (2002) *Proteins*, **46**, 295–307.

Zuker, M. (1989) *Science*, **244**, 48–52.

World Wide Webs cited

AAindex database:	http://www.genome.ad.jp/dbget
AMBER server:	http://www.amber.ucsf.edu/amber/amber.html
B server:	http://www.scripps.edu/~nwhite/B/indexFrames.html
Cambridge Crystallographic Data Centre:	http://www.ccdc.cam.ac.uk/
CambridgeSoft:	http://www.camsoft.com
Cn3D:	http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.html
DNA mfold:	http://bioinfo.math.rpi.edu/~mfold/dna
HyperCube:	http://www.hyper.com
KineMage:	http://orca.st.usm.edu/~rbateman/kinemage/
MDL ISIS Draw:	http://www.mdli.com/dwonload/isisdraw.html
NPS@):	http://npsa-pbil.ibcp.fr
Protein Data Bank:	http://www.rcsb.org/pdb/
RasMol:	http://www.umass.edu/microbio/rasmol/
RNA mfold:	http://bioinfo.math.rpi.edu/~mfold/rna
Swiss-Pdb Viewer:	http://www.expasy.ch/spdbv/mainpage.html
WaveFunction:	http://www.wavefun.com

CHAPTER 10

BIOMACROMOLECULAR INTERACTION

10.1 BIOMACROMOLECULES IN SOLUTION

A solution is a single-phase system containing more than one component, of which biomacromolecules constitutes one component. In an interacting (binding) reaction between a biomacromolecule and its ligand, only three independent variable components are considered if the system is in equilibrium. Specification of the amounts of solvent and any two of the interacting components via the equilibrium relationships, determine the rest (more than three components needed to be specified for a nonequilibrium system). Thus the state of the system in equilibrium is defined by specifying these component amounts (concentrations), temperature and pressure. Most biochemical equilibria involve biomacromolecules, either in reactions with other biomacromolecules or in reactions with small ligands. Almost all of these reactions occur in solution. For a biochemical reaction (in an ideal solution):

$$\mathrm{aA + bB \rightarrow pP + qQ}$$

The reaction quotient, Q is expressed as

$$\mathrm{Q = [P]^p[Q]^q/([A]^a[B]^b)}$$

The driving force of the reaction, which is the free-energy change, ΔG can be written as

$$\Delta \mathrm{G} = \Delta \mathrm{G^o} + \mathrm{RT \ln Q}$$

At equilibrium ΔG = 0, the standard free energy change, ΔG^o is related to the equilibrium constant, K according to

$$\Delta \mathrm{G^o} = -\mathrm{RT \ln \{[P]^p[Q]^q/([A]^a[B]^b)\}_{eq}} = -\mathrm{RT \ln K}$$

where the equilibrium constant $\mathrm{K = \{[P]^p[Q]^q/([A]^a[B]^b)\}_{eq}}$ and therefore

$$\mathrm{K = exp(-\Delta G^o/RT)}$$

The equilibrium constant can now be related to the other thermodynamic states, the standard enthalpy change (ΔH^o) and the standard entropy change (ΔS^o) by

$$-\mathrm{RT \ln K} = \Delta \mathrm{G^o} = \Delta \mathrm{H^o} - \mathrm{T}\Delta \mathrm{S^o}$$

or

$$\mathrm{\ln K} = -\Delta \mathrm{H^o/(RT^2)} + \Delta \mathrm{S^o/R}$$

Experimentally, ΔH^o can be calculated from K as the function of T according to

Biomacromolecules, by C. Stan Tsai
Copyright © 2007 John Wiley & Sons, Inc.

$$\ln K(T) = -\Delta H^{o}/R(T^{-1}) + \text{constant}$$

Thus the thermodynamic variables, ΔG^{o}, ΔH^{o} and ΔS^{o} for a reaction can be obtained by measuring equilibrium concentrations at different temperatures.

Biomacromolecules often carry many charge groups. For example, a B-DNA duplex has two phosphate groups (with a negative charge per phosphate group at physiological pH) every 3.4 Å along the molecule. If this DNA solution contains only Na^+ as the counter ions and ψ as the fraction of Na^+ condensed per charged DNA phosphate, an interaction (binding) of a protein, P (mole/L) displaces n phosphate groups from the DNA, D (expressed as mole phosphate/L), i.e.:

$$D + P \rightleftharpoons DP + n\psi Na^+$$

The equilibrium constant can be written as

$$K = [DP][Na^+]^{n\psi}/([D][P])$$

and the observed binding constant (quotient) as

$$Q = [DP]/([D][P])$$

Thus

$$Q = K\ [Na^+]^{-n\psi}$$

or

$$\log Q = \log K - n\psi \log[Na^+]$$

Experimentally, interaction constants are measured in solutions containing different amounts of Na^+ and a linear relationship between $\log Q$ and $\log[Na^+]$ gives the slope of $-n\psi$ (upper limit for n because $\psi \leq 1$). For double-stranded DNA, $\psi \approx 0.88$. Therefore the number of phosphate groups on the DNA that interacts with the protein (n) can be evaluated.

Most biomacromolecular interactions (bindings) involve the formation of some types of noncovalent bond and some specific region of the macromolecule called the binding site. Determinations of equilibrium binding processes are mostly indirect in that either: i) the fraction of all of the ligand molecules (generally labeled molecules are used) in the system that are bound; or ii) the fraction of binding sites that are occupied, is measured.

A technique commonly employed in binding studies involves equilibrium dialysis in which the concentrations of free and/or bound ligands in equilibrium with the biomacromolecules are measured, where $[A_t] = [A] + [A_b]$ ($[A_t]$, $[A]$, and $[A_b]$ are respectively molar concentrations of the total, free and bound ligands). If the total molar concentration of biomacromolecule is $[P_t]$, the number of moles of A bound per mole of P, ν is given by

$$\nu = [A_b]/[P_t]$$

and

$$\lim_{[A]\to 0} \nu = n,$$

i.e. ν approaches the limit of the number of binding sites (n) if an experiment is conducted at a high enough concentration of A that saturates all the available binding sites. The parameters ν and n can be analyzed to yield binding constants and mechanism of binding. Other related experiments involve gel filtration chromatography or sedimentation.

Another frequently used technique involves the detection of some physical or chemical change in either biomacromolecule or the ligand upon binding, such as differential spectrometry or changes in biological activities. The fractional saturation, θ, which expresses the fraction of sites in biomacromolecule occupied by ligand, A can be written as

$$\theta = \nu/n = \Delta X/(\Delta X)_t$$

where ΔX and $(\Delta X)_t$ are the observed change at a given A and the total change produced when the biomacromolecule is saturated with A. In order to apply this technique to binding studies, the change in parameter X must be linear in ν and it must be the same for each site in a multi-site macromolecule. Since the technique only determines the fraction of total sites occupied, it would not yield information regarding *n*, which has to be determined by other methods. However, if the value of *n* is known, ν can be evaluated from $\nu = n\Delta X/(\Delta X)_t = [A_b]/[P_t]$.

10.2 MULTIPLE EQUILIBRIA

10.2.1 Single-site binding

Most physiological processes result from a ligand (an effector) interaction with a biomacromolecule (Harding and Chowdhry, 2001; Weber, 1992), such as interactions between enzymes and their substrates, between hormones and hormone receptors, between antigens and antibodies, between inducers and DNA, and so on. In addition, there are macromolecule–macromolecule interactions such as between proteins (Golemis, 2002; Kleanthous, 2000), between protein and nucleic acids (Haynes, 1999; Saenger and Heinemann, 1989; Travers, 1993), and between protein and saccharides. A general biomolecular interaction network database (BIND) is available at http://www.bind.ca.

The effecter of small molecular weight is normally referred to as the ligand and the biomacromolecule is known as the receptor. In order to introduce some of formalism and methods of data analysis in binding studies, the simplest case involving a macromolecular receptor (P) having only a single site for binding of a ligand (A) will be considered first. The binding constant (association constant) for the reaction:

$$P + A \rightleftharpoons PA$$

is defined as

$$K = [PA]/[P][A]$$

The average number of moles of bound ligand molecules per biomacromolecular receptor, ν can be represented by

$$\nu = [PA]/([P] + [PA]) = K[P][A]/([P] + K[P][A]) = K[A]/(1 + K[A])$$

The equation describes ν versus [A] as a rectangular hyperbola and the double reciprocal transformation gives a linear relationship:

$$1/\nu = 1 + K^{-1}/A$$

It is noted that the fractional saturation:

$$\theta = \nu = K[A]/(1 + K[A])$$

for the single-site binding in which $n = 1$.

10.2.2 Multiple-site binding: General

The biochemical interaction systems are characterized by general ligand interactions at equilibrium, site–site interactions and cooperativity as well as linkage relationships regarding either (A) different ligands binding to the same macromolecule or (B) the same ligand binding to the different sites of the multi-site macromolecule (Steinhardt and Reynolds, 1969). The multi-site binding is described by multiple equilibrium, which involves a macromolecular receptor, P having *n* number of sites for binding of a ligand, A. Each site has the microscopic ligand association constant, K_i for the i-site:

$$\begin{array}{lll}
P + A \rightleftharpoons PA & K_1 = [PA]/([P][A]) & [PA] = K_1[P][A] \\
PA + A \rightleftharpoons PA_2 & K_2 = [PA_2]/([PA][A]) & [PA_2] = K_1K_2[P][A]^2 \\
\cdots\cdots & \cdots\cdots & \cdots\cdots \\
PA_{i-1} + A \rightleftharpoons PA_i & K_i = [PA_i]/([P_{i-1}][A]) & [PA_i] = K_1K_2 \ldots K_i[P][A]^i \\
\cdots\cdots & \cdots\cdots & \cdots\cdots \\
PA_{n-1} + A \rightleftharpoons PA_n & K_n = [PA_n]/([PA_{n-1}][A]) & [PA_n] = K_1K_2 \ldots K_{n-1}K_n[P][A]^n
\end{array}$$

The total molar concentration of the biomacromolecule, $[P_t]$ is

$$\begin{aligned}
[P_t] &= [P]+[PA]+2[PA_2]+3[PA_3]+\cdots\cdots = \sum_{i=0}^{n}[PA_i] \\
&= [P]\{1+K_1[A]+K_1K_2[A]^2+\cdots\cdots+K_1K_2\cdots\cdots K_{n-1}K_n[A]^n\} = [P]\left\{1+\sum_{i=1}^{n}\left(\prod_{j=1}^{i}K_j\right)\right\}[A]^i
\end{aligned}$$

and the concentration of bound ligand, $[A_b]$ is

$$\begin{aligned}
[A_b] &= [PA]+2[PA_2]+3[PA_3]+\cdots\cdots = \sum_{i=0}^{n}i[PA_i] \\
&= [P]\{K_1[A]+2K_1K_2[A]^2+\cdots\cdots+nK_1K_2\cdots\cdots K_{n-1}K_n[A]^n\} = [P]\left\{\sum_{i=1}^{n}\left(i\prod_{j=1}^{i}K_j\right)\right\}[A]^i
\end{aligned}$$

It is noted that the summation for $[P_t]$ starts at i = 0 to include the unliganded term, [P] and the summation for $[A_b]$ is associated with the numerical coefficients to indicate that each mole of $[PA_i]$ carries i moles of ligand. The equilibrium measurement in binding studies typically yields the average number of moles of ligand bound per mole of biomacromolecule, ν which is given by

$$\nu = [A_b]/[P_t] = \frac{\sum_{i=1}^{n}\left\{\left(\prod_{j=1}^{i}K_j\right)[A]^i\right\}}{1+\sum_{i=1}^{n}\left\{\left(\prod_{j=1}^{i}K_j\right)[A]^i\right\}}$$

This is the general equation for multiple equilibrium (Σ and Π are respective symbols for summation and multiplication) that describes the binding of ligand molecules to the biomacromolecules with multiple sites. However, there are so many adjustable parameters in this general equation, which is more revealing about the binding mechanism if the experimental data can be fitted to simplified models by more restrictive equations with fewer adjustable parameters. These models will be considered.

10.2.3 Multiple-site binding: Equivalent sites

The first case of multiple-site bindings involves a biomacromolecule with n equivalent sites (each of the n sites has the same affinity, K for the ligand) and the binding is noncooperative (each of the n sites is noninteracting/independent). Consideration of the stereochemistry of the bounded A for n equivalent sites suggests that there are

n forms of PA	with $K_1 = nK$
$n(n-1)/(1\cdot 2)$ forms of PA_2	with $K_1K_2 = \{n(n-1)/1\cdot 2\}K^2$
$n(n-1)(n-2)/(1\cdot 2\cdot 3)$ forms of PA_3	with $K_1K_2K_3 = \{n(n-1)(n-2)/(1\cdot 2\cdot 3)\}K^3$
. . . . . .	. . . . . .
$n!/[(n-i)!i!]$ forms of PA_i	with $K_1K_2 \ldots K_i = \{n!/[(n-i)!i!]\}K^i = \Pi K_i$
. . . . . .	. . . . . .
1 form of PA_n	with $K_1K_2 \ldots\ldots K_n = 1K_n$

The number of a particular liganded isomeric forms, PA_i is given by the number of ways in which *n* sites may be divided into i occupied sites and (n – i) vacant sites, such that there are $n!/[(n-i)!i!]$ isomeric forms. Thus it can be shown that (since $\Pi K_I = n!/[(n\text{-}i)!i!]\}K^I$):

$$\nu = \frac{\sum (i\Pi K_j)\}[A]^i}{1+\sum \Pi K_j\}[A]^i} = \frac{\sum i\{n!/[n-i)!i!\}(K[A])^i}{1+\sum \{n!/[n-i)!i!\}(K[A])^i}$$

$$= \frac{nK[A](1+K[A])^{n-1}}{(1+K[A])^n} = \frac{nK[A]}{1+K[A]}$$

This equation is remarkably similar to the equation for single site binding (except the inclusion of n) in that all the n sites on the biomacromolecule are equivalent and independent, therefore the ligand molecule is unable to differentiate one site from the other and experience their effects. In this case, the fraction of all the sites that are occupied:

$$\theta = \nu/n = K[A]/(1 + K[A])$$

and

$$\nu = nK[A]/(1 + K[A])$$

which is identical to that derived from the general equation for multiple equilibrium.

Experimentally, ν are measured by varying [A] and data analysis is performed to evaluate two binding parameters, n and K, often via linear transformation such as the Klotz equation and the Scatchard equation (Table 10.1).

TABLE 10.1 Experimental evaluation of binding parameters

Approach	Direct	Klotz equation	Scatchard equation
Equation,	$\nu = nK[A]/(1 + K[A])$	$\nu^{-1} = n^{-1} + (nK)^{-1}[A]^{-1}$	$\nu/[A] = nK - \nu K$
Function	Hyperbola	Linear	Linear
Plot	ν versus [A]	ν^{-1} *versus* $[A]^{-1}$	ν/[A] versus ν
Evaluation of n	Asymptote = n (est.)	Intercept at $\nu^{-1} = 1/n$ or Extrapolated intercept at $[A]^{-1} = -K$	Intercept at ν = n or Intercept at ν/[A] = nK
Evaluation of K	[A] = 1/K at ν = n/2	Slope = 1/(nK)	Slope = −K

For the case of multiple-site bindings involving a biomacromolecule with n equivalent but interacting sites, the analysis includes an interaction coefficient, h representing the strength of interaction such as:

$$\begin{array}{ll} P + hA \rightleftharpoons PA^h & K_1 = [PA^h]/([P][A]^h) \\ PA + hA \rightleftharpoons PA_2^h & K_2 = [PA_2^h]/([PA^h][A]^h) \\ \ldots\ldots & \ldots\ldots\ldots \\ PA_{n-1}^h + hA \rightleftharpoons PA_n^a & K_n = [PA_n^h]/([PA_{n-1}^h][A]^h) \end{array}$$

The expression for ν yields

$$\nu = nK[A]^h/(1 + K[A]^h)$$

The rearrangement gives

$$\nu/(n - \nu) = K[A]^h$$

This equation, known as the Hill equation, is often written in the logarithmic form:

$$\log \nu/(n - \nu) = \log K + h \cdot \log[A]$$

or

$$\log \theta/(1 - \theta) = \log K + h \cdot \log[A]$$

This gives a linear relationship (plot) of $\log \nu/(n - \nu)$ or $\log \theta/(1 - \theta)$ versus log[A] with slope = h and intercept = log K. The interaction coefficient, h is also called the Hill's coefficient (n_H) and is a measure of the degree of site–site interaction, i.e., $n \geq h \geq 1$. The closer the quantity, h (n_H) approaches the number of sites n, the stronger the interaction, i.e. $h \approx n$ for a biomacromolecule with strongly interacting multiple sites and h = 1 for a macromolecule with noninteracting multiple sites.

10.2.4 Multiple-site binding: Nonequivalent sites

For the case of multiple-site bindings involving a biomacromolecule with n nonequivalent, but not interacting/cooperative sites, the macromolecule is considered to display m classes of independent sites ($n \geq m \geq 1$), each class i ($1 \leq i \leq m$) having n_i sites with an intrinsic binding constant, K_i (Klotz and Hunston, 1971). The total number of sites occupied by a ligand per macromolecule will simply be the sum of the number of sites of each class occupied per macromolecule ([A] will be written simply as A for the clarity here), i.e.:

$$\nu = \sum_{i=1}^{m} \nu_i = \sum [n_i K_i A/(1 + K_i A)]$$

The equation can be rewritten in Klotz form (double reciprocal) as

$$\nu^{-1} = \{\Sigma[n_i KiA/(1 + K_i A)]\}^{-1} = \{\Sigma[n_i K_i/(K_i + A^{-1})]\}^{-1}$$

The summation is carried out from i = 1 to m and $\Sigma n_i = n$ (total number of sites).

The ν^{-1} versus $[A]^{-1}$ plot according to the above equation generally produces a biphasic curve consisting of two approximately linear segments with the following characteristics:

- From lower ν^{-1} and $[A]^{-1}$ segment:
  Intercept: $\lim_{A^{-1} \to 0} \nu^{-1} = \{\Sigma n_i K_i/K_i\}^{-1} = 1/\Sigma n_i = 1/n$

TABLE 10.2 Responses of multiple-site receptors to various binding plots

		n-Equivalent sites		n-Nonequivalent sites
Equation	Plot	Non-interacting	Interacting	
Klotz	ν^{-1} vs A^{-1}	Linear	Concave up/down	Nonlinear
Scatchard	ν/A vs ν	Linear	Concave down/up	Nonlinear
Hill	$\log \nu/(n - \nu)$ vs $\log A$	Linear, $h = 1$	Linear, $1 \leq h \leq n$	Nonlinear

Slope(L): $d\nu^{-1}/dA^{-1} = \{\Sigma[n_iK_i/(K_i + A^{-1})^2]\}/\{\Sigma[n_iK_i/(K_i + A^{-1})]\}^2$ and thus

$$\lim_{A^{-1}\to 0} \nu^{-1}(d\nu^{-1}/dA^{-1}) = \Sigma(n_i/K_i)/(\Sigma n_i)^2 = \Sigma(n_i/K_i)/n^2 \text{ (or } \Sigma(n_iK_i)/n^2\Pi K_i)$$

- From higher ν^{-1} and $[A]^{-1}$ segment:
  Slope(H): $d\nu^{-1}/dA^{-1} = \{\Sigma[n_iK_i/(K_i + A^{-1})^2]\}/\{\Sigma[n_iK_i/(K_i + A^{-1})]\}^2$ and thus

$$\lim_{A^{-1}\to\infty} \nu^{-1} \lim (d\nu^{-1}/dA^{-1}) = \Sigma(n_iK_i)/\{\Sigma(n_iK_i)\}^2 = 1/\Sigma(n_iK_i)$$

Therefore, n, Σn_iK_i (or $\Sigma ni/Ki$) and ΠK_i can be evaluated. An insight into the relative magnitude of the binding constants is feasible if m is small and known.

For the case of multiple-site bindings involving a biomacromolecule with n nonequivalent and interacting/cooperative sites, the general equation for the multiple equilibrium applies and its solution is complex and difficult. However, an intuitive observation of various plots may suggest the likely behavior of multiple-site biomacromolecules (Table 10.2). Both the Klotz and Scatchard plots serve to differentiate noninteracting (linear) versus interacting (nonlinear) n-equivalent sites, while the Hill's plot serves to differentiate between the n-equivalent (linear) versus n-nonequivalent (nonlinear) sites. The interacting/cooperative multiple-site macromolecular receptors are generally oligomeric, consisting of homooligomers or heterooligomers with different conformations. The cooperativity of these oligomeric receptors to the ligand bindings is an important regulatory mechanism for biomacromolecules and is the topic of the next section.

10.3 ALLOSTERISM AND COOPERATIVITY

10.3.1 Models

The ligand-induced conformational changes appear to be an important feature of receptors with regulatory functions. The concepts of cooperativity and allosterism (Koshland *et al.*, 1966; Monad *et al.*, 1965; Richard and Cornish-Bowden, 1987) in proteins have been applied to explain various ligand-receptor interaction phenomena. The observed changes in successive association constants for a multiple ligand binding system suggest the cooperativity in the interaction between binding processes. If binding of a ligand to one site on a biomacromolecule affects the affinity of other sites for the ligand, the binding is said to be cooperative. The interaction that causes an increase in successive association constants is called positive cooperativity, while a decrease in successive association constants is the consequence of negative cooperativity. The cooperativity refers to interaction between binding sites in which the binding of one ligand modifies the ability of a subsequent ligand molecule to bind to its binding site, whereas the allosterism refers to the binding of ligand molecules to different sites.

Because a single binding site cannot generate cooperativity, a cooperative binding system must consist of two or more binding sites on each molecule of biomacromolecules. Although there is no need for such receptors to be oligomeric, nearly all known cases of cooperativity at equilibrium are found in proteins with separate binding sites on different subunits. Furthermore, the cooperativity can be observed with interactions of a single kind of ligand in homotropic (same ligand) interactions or those involving two (or more) different kinds of ligands in heterotropic interactions. Biomacromolecule-ligand interactions are concerned mainly with the homotropic interactions, which will be discussed here (heterotropic interactions in the presence of inhibitor/activator will be considered in the next chapter). Two models, the symmetry model and the sequential model, have been proposed for the cooperativity of homotropic interactions between ligands and proteins. They are also applicable to other macromolecular receptors. The two models differ in that the symmetry model predicts only positive cooperativity whereas both positive and negative cooperativities are possible according to the sequential model. These two models are highlighted below.

- The *symmetry model* of Monad, Wyman and Changeux (Monad *et al.*, 1965):

This model was originally termed the allosteric model. The model (Figure 10.1) is based on three postulates about the structure of an oligomeric protein (allosteric protein) capable of binding ligands (allosteric effectors):

1. Each subunit (protomer) in the protein is capable of existing in either of two conformational states, namely T (tight) and R (relaxed) states.
2. The conformational symmetry is maintained such that all protomers of the protein must be in the same conformation, either all T or all R at any instance. Therefore a protein with n subunits is limited to only two conformational states, T_n and R_n. Thus only a single equilibrium constant $L = [T_n]/[R_n]$ is sufficient to express the equilibrium between them.
3. The association constants, K_T and K_R for binding of ligand S to protomers in the T and R conformations respectively are different, with the ratio $K_T/K_R = C$ that is assumed to favor binding to the R conformation.

The fractional saturation has the following expression:

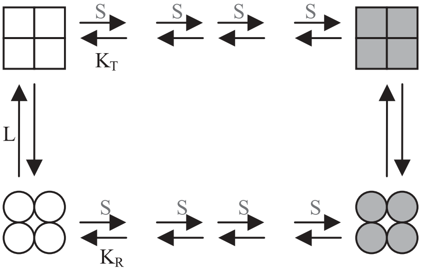


Fig. 10.1 Tetrameric example for the symmetry model
The oligomeric macromolecule is assumed to exist in equilibrium between two conformational states, T (square) and R (circle) with an equilibrium constant, L. The filled squares and circles indicate the liganded (S-bounded) protomers. The association constants for the binding of S to T and R protomers are $K_T = TS_i/TS_{i-1} \cdot S$ and $K_R = RS_i/RS_{i-1} \cdot S$ (*note*: association constants are used in accordance with the practice in molecular interactions) respectively

$$\nu = \frac{\sum_{i=1}^{n}(iRS_i) + \sum_{i=1}^{n}(iTS_i)}{\sum_{i=0}^{n}(RS_i) + \sum_{i=0}^{n}(TS_i)} = \frac{R\cdot\sum_{i=1}^{n}\left[i\cdot n!/\{(n-i)!i!\}(K_RS)^i\right] + T\cdot\sum_{i=1}^{n}\left[i\cdot n!/\{(n-i)!i!\}(K_TS)^i\right]}{R\cdot\sum_{i=0}^{n}\left[n!/\{(n-i)!i!\}(K_RS)^i\right] + T\cdot\sum_{i=0}^{n}\left[i\cdot n!/\{(n-i)!i!\}(K_TS)^i\right]}$$

$$= \frac{\alpha(1+\alpha)^{n-1} + LC\alpha(1+\alpha)^{n-1}}{(1+\alpha)^n + L(1+C\alpha)^n}$$

in which $L = T/S$, $\alpha = SK_R$, and $C\alpha = SK_{T.}$ remembering $\Sigma_{i=0}\,[n!/\{(n-i)!i!\}\alpha^i] = (1+\alpha)^n$.

If $C = 0$, i.e. S only binds to R-state (most allosteric cases), the equation becomes

$$\nu = \alpha(1+\alpha)^{n-1}/\{L + (1+\alpha)^n\}$$

After reciprocal transformation, division and rearrangement:

$$\alpha/\nu - \alpha - 1 = L/\{(1+\alpha)^{n-1}\}$$

A linear relationship of $\log[(\alpha/\nu) - \alpha - 1]$ versus $\log(1+\alpha)$ is obtained by taking logarithms (Horn and Börnig, 1969):

$$\mathrm{Log}(\alpha/\nu - \alpha - 1) = \log L - (n-1)\log(1+\alpha)$$

The slope, $(n - 1)$ provide an estimate of n.

If $C = 0$ and very small L (S binds only to R-state, which is the favored conformation), the above equation is simplified to

$$\nu = \alpha(1+\alpha)^{n-1}/\{L + (1+\alpha)^n = \alpha/(1+\alpha).$$

This expression is equivalent to Michaelis–Menten equation (rate equation for the one-substrate enzymatic reactions), i.e.

$$v/V = K_RS/(1 + K_RS)$$

- The *sequential model* of Koshland, Némethy and Filmer (Koshland *et al.*, 1966):

The sequential model proposes that the conformational stability of each subunit be determined by the conformations of the subunits with which it is in contact. The model (Figure 10.2) is based on three postulates:

1. Each subunit in the protein is capable of existing in either of two conformational states A and B.
2. There is no requirement for conformational symmetry, therefore mixed conformations are permitted. The stability of any particular state is determined by a product of equilibrium constants including a term K_t for each subunit in the B conformation, expressing the free energy of the conformational transition from A to B of an isolated subunit, a term K_{AB} for each contact between a subunit in the A conformation with one in the B conformation, and a term K_{BB} for each contact between a pair of subunits in the B conformation.
3. The ligand binds only to conformation B, with association constant K_S. The binding to conformation A does not exist.
   The derivation of a binding equation generally involves
   - **a)** the designation of the molecular species assumed to be present;
   - **b)** the derivation of the individual molecular parameters for each molecular species; and
   - **c)** the summation of these individual parameters into an overall equation.

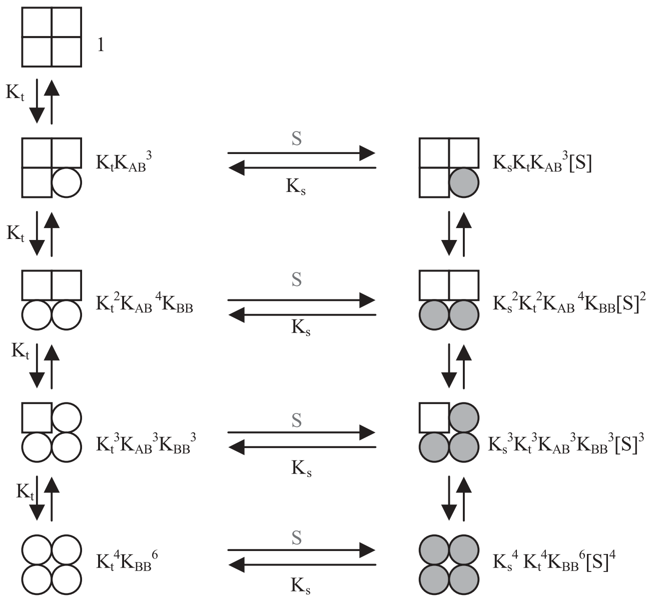


Fig. 10.2 Tetrameric example for the sequential model
The subunits (protomers) of an oligomeric macromolecule are capable of existing in two conformations, A (square) and B (circle). The conformational transition is described by K_t. Two interaction constants associated with subunits are K_{AB} (between subunits A and B) and K_{BB} (between subunits B). The ligand molecule can only bind to the subunit in B conformation with an association constant of K_s. The filled circles indicate the liganded B-subunits. The square arrangement of the tetramer is shown for the clarity of the presentation. However, the preferred tetrahedral geometry of which four subunits are equivalent, is considered here. The individual molecular parameters for each molecular species in the tetrahedral tetramer are given

Because the sequential model deals with subunit interactions explicitly and refers to contacts between subunits, it is necessary to consider the geometry of the molecule, e.g. a tetrameric protein may exists as a square or tetrahedral of which tetrahedral is the preferred arrangement. Treatment of the tetrahedral tetrameric protein gives rise to the following expression:

$$\nu = \frac{K_SK_tK_{AB}{}^3[S] + 3K_S{}^2K_t{}^2K_{AB}{}^4K_{BB}[S]^2 + 3K_S{}^3K_t{}^3K_{AB}{}^3K_{BB}{}^3[S]^3 + K_S{}^4K_t{}^4K_{BB}{}^6[S]^4}{1 + K_SK_tK_{AB}{}^3[S] + K_S{}^2K_t{}^2K_{AB}{}^4K_{BB}[S]^2 + K_S{}^3K_t{}^3K_{AB}{}^3K_{BB}{}^3[S]^3 + K_S{}^4K_t{}^4K_{BB}{}^6[S]^4}$$

Since the product of constants yield another constant ($\Pi K_x = K$), such as $K_sK_tK_{AB}{}^3 = K_1$, $K_sK_tK_{AB}K_{BB} = K_2$ or $K_sK_tK_{AB}'K_{BB}' = K_j$ (where $K_{AB}' = K_{AB}{}^{\pm m}$ and $K_{BB}' = K_{BB}{}^{\pm n}$), the above equation can be rewritten as

$$\nu = \frac{K_1[S] + 3K_1K_2[S]^2 + 3K_1K_2K_3[S]^3 + K_1K_2K_3K_4[S]^4}{1 + K_1[S] + K_1K_2[S]^2 + K_1K_2K_3[S]^3 + K_1K_2K_3[S]^4} = \frac{\sum\{(i\Pi K_j)[S]^i\}}{1 + \sum\{(\Pi K_j)[S]^i\}}$$

This is similar to the general equation for multiple-site equilibrium. Following the previous discussions, therefore in the case of non-cooperativity (non-interacting):

$$\nu = nK[S]/(1 + K[S]).$$

Similarly in the case of infinite cooperativity (maximum interaction, i.e. all intermediate molecular species are negligible):

$$\nu = nK[S]^h/(1 + K[S]^h)$$

where h is the Hill's interaction coefficient that measures the extent of cooperativity among multiple ligand binding sites (Weiss, 1997).

Treatment of other geometries leads to the same expressions for the concentrations, except for the subunit interaction terms, which are different for each geometry.

10.3.2 Diagnostic tests for cooperativity

It is noted that the application of the multiple equilibrium to cooperativity, though fundamentally important, is not practical because it requires knowledge of the number of binding sites and the values of the individual association constants. In practice, assumptions are made to simplify the model and therefore to limit the molecular species being considered to manageable number. Thus diagnostic tests are needed to indicate whether the assumptive model is pertinent and, if not, what molecular species may contribute significantly to the model. It is often convenient to define site equivalency and cooperativity with reference to readily available experimental observations. Table 10.3 lists some of the tests.

Operationally, the Hill equation (or the Hill plot) provides a good measure of cooperativity. The Hill coefficient, h (or the slope of the Hill plot) is larger than 1 for positive cooperativity, less than 1 for negative cooperativity and equal to 1 for noncooperativity. Thus the ligand-biomacromolecule interaction is positively cooperative, negatively cooperative or noncooperative according to the sign of (h − 1). In the following sections, examples of molecular interactions involving proteins, nucleic acids and glycans will be discussed.

TABLE 10.3 Diagnostic tests for cooperativity of oligomeric biomacromolecules

Test	Analysis	Observation	Conclusion
Direct binding	ν vs S plot	Hyperbolic	Equivalent, non-cooperative
		Sigmoid	Positive cooperativity
		Hyperbolic, reduced asymptote	Negative cooperativity
Klotz equation	ν^{-1} vs A^{-1} plot	Straight line	Non-cooperativity
		Concave-up at low A^{-1}	Positive cooperativity
		Concave-down at low A^{-1}	Negative cooperativity
Scatchard equation	ν/A vs ν	Straight line	Non-cooperativity
		Concave-up at low ν	Negative cooperativity
		Concave-down at low ν	Positive cooperativity
Hill equation	$\log \nu/(n - \nu)$ vs $\log A$	Line: slope (h) = 1	Non-cooperativity
		slope (h) > 1	Positive cooperativity
		slope (h) < 1	Negative cooperativity
		Curved	Non-equivalent sites
Cooperative index[a]	$R_s = S_{90}/S_{10}$	$R_s = 81$	Non-cooperativity
		$R_s > 81$	Negative cooperativity
		$R_s < 81$	Positive cooperativity

Note: [a] The cooperative index (R_s) refers to the ratio of two ligand concentrations to produce 90% (S_{90}) and 10% (S_{10}) fractional saturation (Koshland *et al.*, 1965).

10.4 SPECIFICITY AND DIVERSITY OF ANTIBODY-ANTIGEN INTERACTIONS

Protein-ligand interactions are important phenomena that touch upon every facet of biological functions. These include such examples as enzyme-substrate interactions in biochemical transformations, transducer-membrane interactions in signal transduction, protein-nucleic acid interactions in genetic transmission, protein-carbohydrate interactions in cell adhesion as well as protein-protein interactions in biochemical regulations and defense (immune response). Databases of interacting proteins are available respectively at DIP (http://dip.doe-mbi.ucla.edu) and IntAct project of EBI (http://ebi.ac.uk/intact).

10.4.1 Structure of antibody

The diversity of the immune response is impressive in that any foreign substances (antigens), small molecules or macromolecules, can elicit productions of appropriate proteins (antibodies) in response to diverse structures of antigens. In addition, the antibody response shows remarkable specificity. The specificity of the antibody-antigen interaction might be regarded as a model for molecular recognition.

Antibodies belong to the class of serum glycoproteins called immunoglobulins (Igs), which are made in all vertebrates as part of the immune response to antigenic challenge by foreign substances (van Oss and van Regenmortel, 1994). Immunoglobulins are Y-shaped proteins made up of two identical light (L) polypeptide chains (Mol. wt. = ~25 kDa) and two identical heavy (H) chains (mol. wt. = ~50 kDa) held together by disulfide bonds. Immunoglobulins are divided into several major classes or isotypes characterized by their heavy (H) chain type (Table 10.4). They all contain carbohydrates, largely D-hexose and D-hexosamine but also sialic acid and L-fucose, covalently attached to protein moiety.

The immunoglobulin G (IgG) class is the most abundant in normal serum and the most commonly observed class in human myeloma Igs. Its diagrammatic structure, including homology regions, is shown in Figure 10.3. Immunoglobulins (antibodies) are oligomeric proteins consisting of globular subunits arranged in pairs with subunits of the light (L) chain of approximately 220 amino acids and the heavy (H) chain of 450–575 amino acids. The light chains contain two immunoglobulin domains; the N-terminal domain is variable, i.e. it varies from antibody to antibody, and the C-terminal domain is constant, i.e. it is the same in light chains of the same type. The heavy chains are made up of an N-terminal variable (V) domain and three or four constant (C) domains. The anti-

TABLE 10.4 Properties of human immunoglobulins

	IgG	IgA	IgM	IgD	IgE
H chain class	γ	α	μ	δ	ε
H chain subclass	$\gamma1, \gamma2, \gamma3, \gamma4$	$\alpha1, \alpha2$	$\mu1, \mu2$		
L chain type	κ and λ	κ and λ	κ and λ	κ and λ	κ and λ
Molecular formula	γ_2L_2	$(\alpha_2L_2)_2/(\alpha_2L_2)_2J$	$(\mu_2L_2)_5J$	δ_2L_2	ε_2L_2
Approx. M_r (kDa)	150	160/400	900	180	190
Serum conc. (mg/dL)	1000	200	120	3	0.05
Stability, $t_{1/2}$ (days)	23	6	5	2–8	1–5

Note: The J (joining) chain is a small glycoprotein (M_r ~ 15 000) with an unusually high content of Asp and Glu.

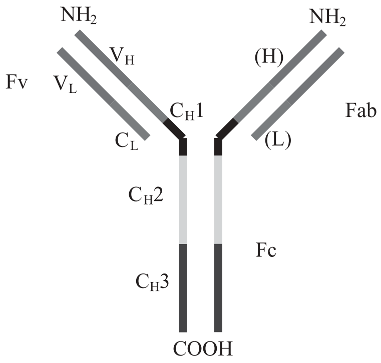


Fig. 10.3 Sketch of IgG molecule
The light (L) chains (mol. wt. ~25 000) are divided into two homology regions, V_L and C_L in which V and C are variable sequence and constant (conserve) sequences. The heavy (H) chains (mol. wt. ~50 000) are divided into four homology regions, V_H, C_H1, C_H2, and C_H3 from the N-terminus toward the C-terminus. The C_H1 and C_H2 are connected by the flexible 'hinge' region that is susceptible to proteolytic cleavage by papain, pepsin and trypsin. The H and L chains are covalently linked by —S-S— bonds. The antigen-binding site is made from the combination of the variable parts (V_H and V_L) of the H and L chains known as Fv fragment. The N segments from the hinge is known as Fab fragment (Pepsin digestion of Ig yields Fab_2) and the C segment from the hinge is known as Fc fragment (papain digestion of Ig yields Fab + Fc). IgM and IgA molecules differ in the Fc fraction

body fragment containing the associated variable domains of the light chain (V_L) and of the heavy chain (V_H) is called the Fv; the fragment containing the entire light chain and the V_H and the first constant domain (C_H1) of the heavy chain is called the Fab. The fragment Fc consists of the second and third constant domains (C_H2 and C_H3) of the two heavy chains.

The constant regions (C_H1, C_H2, C_H3 and C_L) are homologous to each other and non-homologous to V_H and V_L. The N-terminal variable regions, V_H and V_L are highly homologous to each other. The homology regions of Igs fold into independent compact units of three-dimensional structure (immunoglobulin fold). The immunoglobulin fold consists of two twisted, stacked β-pleated sheets that surround an internal volume tightly packed with hydrophobic side chains. These two β-sheets are covalently linked by an intrachain disulfide bridge in the inner volume, in a direction approximately perpendicular to the plane of the sheets. About 50% of the amino acid residues of the subunit forms part of the β-sheets having highly conserved sequences in different immunoglobulins. The V subunits have an extra length of polypeptide chain that form the two-stranded loop. The major amino acid differences between Igs of the same class and of the same animal species occur in the loops connecting the β-pleated sheet strands. Useful information concerning Igs can be obtained at IMGT (http://imgt.cines.fr) and KABAT (http://immuno.bme.nwu.edu/).

A large number of noncovalent interactions stabilize the arrangement of the different homology subunits in Ig molecules. The interaction occurs between adjacent subunits of the same chain (*cis* interactions) and between subunits in different chains (*trans* interactions). *Trans* interactions are in general extensive and stabilize the structural domains formed by pairs of homology subunits. Conversely *cis* interactions involve a small number of contact residues and the subunit arrangements stabilized by these interactions are generally flexible. In Igs the homology subunits are found associated in pairs through *trans* interactions. With the exception of C_H2 region, the structural association between subunits is close and involves a large area of contact. The interactions between consecutive subunits of the same chain in Ig molecules are limited and involve a small number of contacts. The different homology subunits arose during evolution by a mechanism of gene duplication and diversification. This mechanism provided the structural basis for the different functions of antibody molecules. The evolutionary mechanism did not alter the overall folding pattern of the subunits.

The L chains of IgG can be antigenically classified into two isotypes (or classes) called κ and λ, each characterized by a unique sequence in their C-terminal regions. IgM, IgA, IgD and IgE possess similar κ and λ light chains but their H chains (called μ, α, δ and ε respectively) are different and are specific to each class. Furthermore, the sequence similarity in L chains may be grouped into subclass within a given class. All chains within a subclass are similar in sequence, except at certain positions within V_L, where extreme variability is observed. These hypervariable sequences constitute the regions of the L-chain structure that come in contact with antigens, so that the presence of different sequence in these regions will result in different antibody specificities (Wu and Kabat, 1970). Similarly, the variable region of the H chain occurs at the N-terminal end of the molecule, is approximately 110 amino acid residues long, and also contains hypervariable regions (Kehoe and Capra, 1971). The larger structural differences are those arising from the existence of deletions and insertions in these regions.

Each antibody-forming cell (B-lymphocyte or its fully differentiated progeny, the plasma cell) is committed to the production of one type of antibody molecule. During the immune response, those B cells (called B cells because they mature in the bone marrow) recognizing a particular antigen, in collaboration with other immunoregulatory cells, are stimulated to replicate and differentiate into antibody-secreting cells. Since most macromolecules have a number of distinct antigenic determinants, they stimulate the expansion of many B-cell clones. Furthermore, a single antigenic determinant can stimulate multiple B-cell clones synthesizing antibodies with slightly different specificities. The net result is that immunization with most antigen results in a polyclonal response and the accumulation of many different antibodies in the serum. Conventional antisera are therefore very heterogeneous. Monoclonal antibodies offered convenience and reliability to produce large quantities of homogeneous antibodies to a diverse array of antigens for structural studies. The three-dimensional structure of antibodies has been investigated and reviewed (Amzel and Poljak, 1979; Davies and Metzer, 1983; Alzari *et al.*, 1988).

The concept of preparing monoclonal antibodies is to isolate a B cell (lymphocyte), which produces the wanted specific antibody (but would not grow in culture) and fuses it with a multiple myeloma cancer cell, which does not produce the wanted antibody but does grow easily. This produces an immortalized B cell, which secretes the wanted antibody. The technique is to immunize mice with an antigen and a few weeks later to prepare B cells from the spleen, which is a secondary lymphoid organ. There is a mixed population of B cells producing different antibodies but including some specific for the injected antigen. They are fused with myeloma cells. This gives fused cells called hybridoma cells that were selected by growing the mixture in a selective medium, which does not allow unfused cells to grow and divide. The hybridomas are collected and inoculated into small wells arranged in rows in a small volume of culture fluid in each well. After suitable growth, the wells are tested for the presence of antibody reacting to the antigen of interest. Once identified, the clone (a clone is a collection of identical cells arising from the multiplication of a single cell) can be grown in unlimited quantities, producing the monoclonal antibody also in unlimited amount. The cell culture can be stored indefinitely in liquid nitrogen and grown whenever more of the antibody is wanted. A monoclonal antibody will bind to only one determinant on a monomeric antigen such as a protein.

Monoclonal antibodies can also be labeled by fluorescent compounds and used to identify specific proteins and their binding studies.

10.4.2 Antibody-antigen complex

The combining site of antibodies is formed almost entirely by six polypeptide segments; three each from the light and heavy chain variable domains. They occur at the ends of the molecule (Y-arms or Fab arms), fully exposed to solvent. These segments display variability in sequence as well as in number of residues, and it is this variability that provides the basis of the diversity in the binding characteristics of the different antibodies. These six hypervariable segments are also referred to as the complementarity determining regions or CDRs, which dictate the conformations of the combining sites. The versatility of the immune response in recognizing the universe of possible antigens is a consequence of the combinatorial association of multiple CDR loops. The amino acid sequences of these segments, unique to each different Igs, determine the specificities of antibody molecules. The segments of polypeptide chain connecting the globular Ig domains show different degree of flexibility. This flexibility of antibodies could be important for the mechanism of antigen binding and the ensuing activation of secondary or effector functions involving domains of the structure removed from the antigen-combining site. For the antigen and effector bindings, there is an unusual degree of molecular flexibility found in antibodies. First, there is a flexible hinge region at the fork of the Fab arms and the Fab arm flexibility about the hinge (Y to T shape changes) should allow variable reach for the antibody and therefore bivalent recognition of differently spaced antigens. Second, Fc wagging may be crucial in allowing effector molecule binding. The flexibility in IgG molecule includes Fab arm rotation, Fab elbow bend, Fab arm wagging and Fc wagging (Burton, 1990).

An unusual feature of the antibody combining site is the abundance of aromatic amino acids, particularly Tyr. Aromatic side chains present large surface areas for interaction with antigen, and because they are relatively rigid, lose little conformational entropy upon complex formation. Tyrosine can also interact with polar functionalities on an antigen via hydrogen bonding through its phenolic group. Two other amino acids, Asn and His, are also more likely to be found in the CDR's than in Ig framework. Asparagine probably serves a predominantly structural role, stabilizing the binding site through hydrogen bonding to other side chains and to the protein backbone. Histidine, on the other hand, is a particularly versatile residue. It can contribute to the recognition of a wide range of antigens as a consequence of its aromaticity and nucleophilicity, as well as acidity–basicity.

The binding to the antigen is by noncovalent bonds. Thus an antibody has two identical antigen-binding sites, which means that it can cross-link antigen molecules. The presence of the hinge region increases the cross-linking ability. A given antibody binds to only a small part of the antigen (in the case of protein antigen only a few amino acids). The specific part of the antigen, which is recognized by an antibody, is called an epitope (a distinguishable antigenic site). Thus a protein antigen provokes the production of a number of different antibodies, each combining with a specific epitope and each produced by a different clone of B cells. The structure-infered antigenic epitopes of known complexes can be obtained at Epitome (http://www.rostlab.org/services/epitome)

Structural investigations of monoclonal antibodies complexed with protein antigens such as lysozyme and influenza virus neuraminidase have been investigated (Davies *et al.*, 1990). Two macromolecules in solution have to overcome large entropic barriers before they can form a tight association. There is the loss of the entropy of free rotation and translation of the separate molecules, as well as the loss of conformational entropy of mobile segments and of side chains upon binding. However, entropy is gained when water mol-

TABLE 10.5 Changes in surface areas and contacts in antibody-antigen complexes

Antigen:	Lysozyme	Lysozyme	Lysozyme	Neuraminidase
in Free				
M.W.	14 000	14 000	14 000	50 000
Surface area, Å^2	5 436	5 414	5 564	14 648
in Complex, buried				
Surface area, Å^2	750	774	680	879
%	14	14	12	6
Residues	24	27	27	32
in Complex, contact				
Residues	14	15	15	21
Monoclonal Antibody:	HyHEL-5	HyHEL-10	D1.3	NC41
in Complex, buried				
Surface area, Å^2	746	721	690	886
Residues	28	30	22	32
in Complex, contact				
Residues	17	19	14	21
CDRs	6	6	6	5
Complex				
vdW contacts	74	111	75	108
Hydrogen bonds	10	14	15	23
Ion pairs	3	1	0	1
Reference	a	b	c	d

Notes: Lysozyme is a small 14.6 kDa glycosidase with specificity for hexasaccharide having alternating β-1,4-linked *N*-acetyl glucosamine and *N*-acetylmuramic acid. The intact neuraminidase is a tetramer of molecular weitht 240 000 (monomeric subunit is given). Each subunit of the tetramer is folded into an unusual array of six β sheets, with each sheet having four antiparallel strands arranged consecutively. The enzyme catalyzes the cleavage of terminal sialic acid from adjacent sugar residues. Neuraminidase and hemagglutinin are the two glycoproteins attach to the influenza viral membrane. They make up the antigenic surface of the virus, and changes in these proteins form the basis for evasion of immune recognition from previous infections. Data extracted from Davies *et al.* (1990). Individual references are a) Sheriff *et al.* (1987); b) Padlan *et al.* (1989); c) Amit *et al.* (1986); d) Tulip *et al.* (1989).

ecules are displaced from the surfaces upon the formation of the new antibody-antigen interface. It appears that water molecules are almost totally excluded from the interface by the close shape complementarity between antibodies and antigens. Enthalpic contributions arise from van der Waals interactions, together with the more specific hydrogen bonds and salt bridges (interactions between charged moieties). All of these interactions have been observed in the complexes. For protein antigens, the size of the antibody-combining site becomes the determining factor in antibody-antigen interactions (Table 10.5). Any hypervariable loop can be important. The light chain contributes from 41 to 44% of the surface area, and the heavy chain contributes form 56 to 59% of surface area in the complex formation.

The specificities of the antibody-antigen interactions are believed to result from the complementarity of their interacting surfaces. Indeed, there is remarkable shape complementarity; i.e. depressions on one surface are filled by protuberances from the other, leaving no holes large enough for even water molecules. A particular relevance of the antibody-antigen interaction is the directionality of the hydrogen bonds, necessitating a hydrogen bond acceptor within a certain distance and with a certain fixed angle of the hydrogen bond donor in order to form a strong bond. On average, the antibody-antigen interfaces have about a dozen hydrogen bonds per 200 Å as compared to one per 200 Å for the interfaces of oligomeric proteins (Janin *et al.*, 1988). This difference occurs presumably because the protein surface that is combining with the antibody is one that is normally exposed to solvent and therefore contains a significant content of polar residues.

10.5 COMPLEMENTARITY IN NUCLEIC ACID INTERACTIONS

10.5.1 DNA-protein interaction

Complementarity is the basis for biomacromolecular interactions. The tremendous efficiency displayed by enzyme catalyses, the uncompromising fidelity displayed by the recognition of regulatory effectors for DNA and RNA, the precise specificity displayed by receptors for their cognate ligand and antibody for particular antigens, are in large part due to the molecular complementarity between the two interacting molecules. The interaction of DNA with protein is one of the most active areas of research in structural biology because of its direct importance to our understanding of cellular control mechanisms (Travers, 1993). Noncovalent interactions between DNA and proteins are characterized most notably by their great stability (equilibrium constant, $K_a \approx 10^9 - 10^{12}\,M^{-1}$) and exquisite specificity. A number of forces contribute to the exceptionally strong and specific binding. Some of these are locally in distinct regions of the protein-DNA interface, whereas others exert a more global influence on complexation (Steitz, 1993; Lilley, 1995). Local forces (direct readout) are responsible for many of the direct protein-DNA contacts that include hydrogen bonds, ionic interactions and van der Waal interactions. Global forces (indirect readout) include long-range electrostatic steering, DNA-induced protein folding and conformational changes, protein-induced DNA distortion such as bending or twisting (Kim *et al.*, 1993) and cooperativity gained through simultaneous DNA recognition by multiple protein modules.

Direct hydrogen bond formation and van der Waals interaction between protein and DNA atoms provide the major forces for sequence-specific recognition of DNA by protein. Hydrogen bonds, which make up a large fraction of the total number of contacts in a nucleic acid-protein complex, are formed between an electron-rich acceptor atom, A and a proton attached to an electronegative donor atom, D, with the A–D bond distance of generally 3 Å or slightly less. Although bonding is stronger when all of the atoms in the A–H–D are arranged linearly, there is good reason to believe that hydrogen bonds with bond angle of 135° or less can make important contributions to nucleic acid-protein complexes. Several types of hydrogen bonding contacts to the edges of DNA/RNA bases are commonly observed, including direct hydrogen bonds to amino acid side chains, direct hydrogen bonds to the protein main chain amides and water mediated contact to side chains or main chain of proteins. The nonpolar interactions contribute prominently to the stability of nucleic acid-protein complexes. While hydrogen bonds require a fairly precise geometric alignment of the interacting partners, nonpolar contacts can take place at any relative orientation that is sterically accessible. This high degree of orientational freedom may facilitate nonpolar interactions in nucleic acid-protein interfaces. Formation of a sequence-specific DNA-protein complex is typically accompanied by the burial of 2000–4000 Å of the nonpolar surface area. Another major type of local interaction between nucleic acids and proteins is the ionic interaction (salt bridge), namely the coulombic interaction of atoms having opposite formal charge. Approximately half (~45%) of the total solvent accessible surface area in B-DNA belong to the phosphate groups of the backbone. The favorable entropy associated with the release of small counterions from the DNA backbone upon complexation is known as polyelectrolyte effect. Thermodynamic data on nucleic acid-protein interactions can be obtained from ProNIT (http://gibk26.bse.kyutech.ac.jp/jouhou/pronit/pronit.html).

The size (diameter and typical lengths) of α-helix are almost perfectly suited to fit into the major grooves of B-form DNA (Warren and Kim, 1978). The most prominent features in the protein-nucleic acid complexes observed to date involve complementarity at the level of secondary structure. A large number of transcriptional regulators interact with DNA primarily via the insertion of an α-helix into the major groove (Brennan and Mattews, 1989; Harrison and Aggarwal, 1990). There is also growing evidence for the participation of β-structures in genetic regulation (Phillips, 1991; Rice *et al.*, 1996). A rudimentary structural complementarity of polypeptide secondary structures to the helical structures of nucleic acids underlies many of the observed modes for protein-nucleic acid interactions.

The proteins bind primarily to the major groove of DNA and to lesser extent to the minor groove of DNA. The arrangement of hydrogen bond donors and hydrophobic groups on the DNA as a function of sequence provides a means for recognition of the sequence of DNA. The bendability or deformability of DNA as a function of the sequence also provides a means for sequence-specific recognition of DNA, because it allows some sequences to take up a particular structure required for binding to protein at a lower free energy cost than other sequences. The phosphate groups have been found to play vital roles in the stabilization of protein-DNA complexes through charge interactions and hydrogen bond formation, though their role in sequence-specific recognition is unclear. Several buried water molecules have been found at the protein-nucleic acid interface. These waters apparently play a role in the stability and specificity of certain DNA-protein complexes. The current opinion is that proteins recognize DNA sequences through direct hydrogen bonds, nonpolar contacts, indirect structural effects and water-mediated interactions (Shakked *et al.*, 1994).

The expression of the genetic information is controlled predominantly through the interaction of regulatory proteins with DNA. Regulatory proteins must have the ability to bind a short unique base sequence specifically (Pabo and Sauer, 1984). Other proteins involve in the structural organization of DNA, such as histones, bind DNA in a sequence independent fashion. Specific interactions are usually defined through hydrogen bonding interactions and salt bridges between functional groups of amino acid side chains or the peptide bonds and groups on the bases in the major or minor grooves. The phosphate backbone would, in general, provide a sequence-independent protein binding site (Schleif, 1988; Pabo and Sauer, 1992). Their contributions will be considered.

10.5.1.1 Helices in recognition. Most of the well-characterized families of DNA-binding proteins use α-helices (sometime termed the recognition helix) to make complementary contacts in the major groove. Although β-sheets or regions (generally with two-fold symmetry axis) of polypeptide chain can also make contact, α-helices are used much more frequently. However, there is no evidence that an isolated helix from any of the known motifs (e.g. helix-turn-helix, homeodomain, zinc finger, leucine zipper) can bind DNA in a site-specific fashion, neither is an isolated helical peptide containing a single α-helix capable of site-specific DNA-binding. It appears, therefore, that the overall binding specificity results from a set of interactions, with some contacts from a recognition helix and some from surrounding regions of the protein. Furthermore, the orientations of the α-helices are different among the different DNA-binding proteins, even within a given family. The main focus is that the overall shape and dimension of an α-helix allow it to fit into the major groove in a number of related, but significantly different ways. Some helices lie in the middle of the major groove and have the axis of the α-helix approximately tangent to the local direction of the major groove, i.e. at an angle of 32° with respect to the plane that is perpendicular to the helical axis of the B-DNA. Other α-helices are tipped at different angles, varying at least 15° in each direction, and some are arranged so that only the N-terminal portion of the α-helix fits completely into the major groove. It appears that surrounding regions of the protein, in addition to the sequences of the recognition helices, help to determine how these α-helices are positioned in the major groove of the DNA duplex.

10.5.1.2 Bases in the major and minor grooves are available for specific protein recognition. Specificity in DNA-protein interactions comes from protein recognition of the linear order of base pairs (bps) by hydrogen bond and salt bridge contacts through the major and minor grooves. These groups in the grooves present hydrogen bonding donor and acceptor sites that can participate in hydrogen bond formation with sites on one or more amino acids in protein. There are one hydrogen bond donor and two acceptor groups on the major groove surface of all four dinucleotide pairs. In addition, there is a methyl group at the C5 position of thymine that can participate in van der Waal interactions. Within the minor groove, there are two hydrogen bond acceptor sites in all four dinucleotide pairs, where the N2 position of G in the C·G/G·C dinucleotide pairs provides a hydrogen bond donor at the center of the minor groove. The unique arrangement of hydrogen bond donor and acceptor sites for each dinucleotide within the major or minor groove must provide the specificity utilized by proteins to discriminate specific region of DNA sequence. For example, Gln or Asn can form two hydrogen bonds with adenine in the major groove of an A·T or T·A pair (Seeman *et al.*, 1976). Similarly the guanidinium nitrogens of Arg can form hydrogen bonds with N7 and O6 positions of guanine. In the minor groove, Asn and Gln can form two hydrogen bonds with the N3 and N2 position of guanine.

Proteins can also recognize particular regions of DNA through an indirect readout mechanism in which contacts are not made with the bases in the major or minor grooves, but with phosphate groups and sugar residues. There is nonuniformity in the structure of the Watson–Crick double helix, since the primary DNA sequence defines a precise shape for the double helix. Primary sequence dictates local twist angles, the specific tilt and roll of bases and bends in DNA, as well as the widths of the major and minor grooves. These structural parameters will precisely position in space the hydrogen bonding donor and acceptor sites in the bases and the phosphate backbone. Most DNA binding proteins are designed to recognize a particular shape or flexibility of the double helix in addition to a direct readout of individual bases in the recognition site.

10.5.1.3 The sugar-phosphate backbone is available for nonspecific protein recognition. The phosphate backbone has a relatively uniform shape and negative charge available for nonspecific recognition by DNA binding proteins. DNA polymerases constitute one class of proteins that must interact with DNA in a sequence-independent fashion. Polymerases must traverse the entire chromosome, contacting every individual bp. DNase I provides another example of nonspecific DNA-protein interactions. The nuclease cuts DNA in a sequence-independent fashion. In a cocrystal between DNA and DNase I, contacts include one stacking interaction between arginine and oxygen atoms of pyrimidiine (Suck *et al.*, 1988). However, the majority of the hydrogen bonding contacts are made between 10 other amino acids and the phosphate backbone.

10.5.1.4 Dimeric binding sites in DNA. Many binding sites for regulatory proteins are dimeric, and are recognized by multimeric proteins, usually homotypic dimers or tetramers. A protein must select its small binding site, in which the contacts that provide the specificity are buried within the major groove, from a large amount of nonspecific sequence. The affinity of a protein for its DNA binding site is a function of the number and strength of electrostatic and hydrophobic interactions between the protein and the DNA. Dimeric binding sites can be used to vary the binding affinity between a protein and multiple binding sites in DNA. Groups of genes controlled by the same regulatory protein but located at different positions on the chromosome are called regulons. Sequence deviation from a preferred consensus DNA binding site creates operators with reduced binding affinity. The dimeric operator, in which the sequence of the two individual recognition sites can vary, allows variation of the affinity of the repressor for DNA. These changes in the affinity of the operator with the protein allow the differences in the timing and level of expression of the gene. Variation in the affinity of a repressor for a DNA binding site can also be accomplished by utilizing heterotypic dimers, each with differential inherent affinities for a single binding site.

The principles guiding site-specific recognition in the DNA-protein interactions can be generalized as:

- Site-specific recognition always involves a set of contacts with the bases and the DNA backbone.
- Hydrogen bonding is critical for recognition, although hydrophobic interaction may also be involved.
- Side chains of the protein make most of the critical contacts.
- There is no simple one-to-one correspondence between the side chains and the bases of contacts.
- The folding and docking of the protein molecule may help steer any particular side chain in the site-specific contact.
- Most of the base contacts are in the major groove. Contacts with purines seem to be especially important.
- Most of the major motifs contain an α-helical region that fits into the major groove of B-DNA.
- Nonspecific recognitions with the DNA backbone usually involve hydrogen bonds and/or salt bridges with the phosphodiester oxygen's.
- Multiple DNA-binding domains usually are required for site-specific recognition. The same motif may be used more than once. Different motifs may also be used in the same complex.
- Recognition is a detailed structural process and sequence-dependent aspects of the DNA structure are important.

Three major motifs that are utilized by eukaryotic regulatory proteins to bind specific DNA sequences (Johnson and McKnight, 1989) are:

Helix-turn-helix. Some transcriptional regulatory proteins that recognize specific DNA sequences are known to bind DNA as dimers, and their binding sites on DNA possess dyad symmetry. These proteins share a distinctive succession of two α-helices separated by a relatively sharp β turn termed the helix-turn-helix motif (Steitz *et al.*, 1982; Harrison and Aggarwal, 1990). They have several criteria:

- The distances between the helices are approximately equivalent and are matched to the distance separating successive major grooves on a ring face of B-DNA.

- The helices are aligned in an antiparallel conformation such that their N-to-C dipole orientations correspond to those of DNA sequences on each half of their dyad-symmetric recognition sites.
- One helix (referred to as the recognition helix owing to its contribution to the atomic contacts with DNA) is locked into place via hydrophobic interactions with another helix that sits above it relative to the putative interface with DNA.

The β turn that occurs between the two helices begins most frequently with Gly and is invariably followed by a residue bearing a hydrophobic side chain. Residues four and seven of the recognition helix are almost always occupied by hydrophobic amino acids and likewise the fourth residue of the other helix. The sequence similarity is noted between the helix-turn-helix motif and homeobox elements.

Zinc finger. This motif is characterized by the repetitive ordered occurrence of Cys_2 and His_2/Cys_2 in the amino acid sequences (C_2—H_2 and C_2—C_2 zinc fingers respectively). Each repeat in the C_2—H_2 zinc finger motif sequesters a single zinc ion via tetrahedral coordination with the spatially conserved Cys and His residues. The amino acid residues intervening between the Cys and His would loop out in a manner facilitating direct interaction with DNA, forming a distinct DNA interaction module termed the zinc finger (Fairall *et al.*, 1986). The C_2—H_2 zinc finger motif found in the transcription factor and gene regulatory proteins, is defined by four metal ligands with a sequence, Cys-$X_{2\text{-}5}$-Cys-$X_{12/13}$-His-$X_{2\text{-}5}$-His and three conserved hydrophobic residues. Each 30-residue motif folds to form an independent domain with a single zinc ion tetrahedrally coordinated between an irregular β-sheet and a short α-helix, which is located between the metal ligands. The C_2—C_2 zinc finger motif found in steroid hormone receptors is absent in the conserved hydrophobic residues (within the motif), with a somewhat larger spacing between the two repeat motifs (15 residues in C_2—C_2 instead of 4–8 residues in C_2—H_2). The antiparallel β-sheet is absent and α-helix extends beyond the zinc ligands in the C_2—C_2 motif (Schwabe and Rhodes, 1991). The two zinc-binding C_2—C_2 motifs in the receptors are folded together to form a single domain with conserved amino acids between the two helices of the motifs.

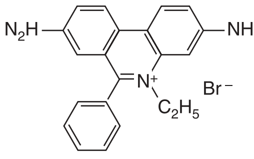
Leucine zipper. The amino acid sequence similarities between the products of several oncogenes and yeast regulatory protein indicate that:

a) 35-amino acid region of the DNA-binding domain contains a strict heptad repeat of Leu residues; and
b) an exceedingly high density of oppositely charged amino acids juxtaposed in a manner suitable for intrahelical ion pairing.

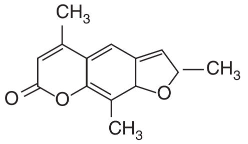
Thus the leucines extending from the helix of one polypeptide chain would interdigitate intimately with those of the analogous helix of a second polypeptide chain, forming an interlock termed the leucine zipper (Landschulz *et al.*, 1988).

10.5.2 Binding of intercalation agent to supercoiled DNA

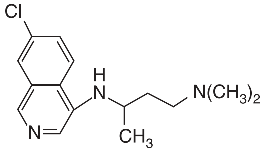
Certain planar aromatic chemicals (drugs) bind to DNA by intercalation. In this process, the intercalation compound inserts between two stacked bps of DNA and positions itself within the center of the DNA double helix. The intercalated molecule is stabilized by hydrophobic stacking interactions with adjacent bps. The binding by intercalation results in several changes in the shape and flexibility of DNA. The binding of an intercalating drug necessarily results in a lengthening of the DNA helix, since two adjacent base pairs must physically separate to accommodate the intercalated molecule.



Ethidium bromide



4,5′,8-Trimethylpsoralen



Chloroquine

Ethidium bromide binds tightly to DNA when irradiated with 360 nm UV light, bound ethidium bromide fluoresces bright red and is used to visualize DNA in agarose/acrylamide gels or CsCl density gradient. 4,5′,8-Trimethylpsoralen and psoralen families bind to DNA and form covalent linkages with pyrimidine bases. They are permeable to both prokaryotic and eukaryotic membranes and have been used for *in vivo* studies of DNA structure. Chloroquine is used to unwind DNA and relax supercoils for topological analysis on agarose gels.

The unwinding from the binding of intercalating drugs relaxes negative supercoils. The unwindings for ethidium bromide and psoralen are about 26° and 28° per intercalated molecule respectively. It takes 12.8 intercalated molecules to unwind the helix by 360° or remove one turn of the DNA double helix with an unwinding angle of 28°. Unwinding from intercalation lowers the value of L. Therefore as an increasing number of molecules intercalate into DNA, the value of L decreases. Since supercoiled DNA contains a deficiency of helical turns, a reduction in L_0 will reduce the level of negative supercoiling. At some point ($L = L_0$) the DNA will become relaxed. In addition for a drug with a high binding constants, as the level of bound drug continues to increase, eventually L will be higher than L_0 and the DNA will become positively supercoiled.

10.5.3 RNA-protein interaction

Specific interactions between RNA and proteins are fundamental to many cellular processes, including the assembly and function of ribonucleoprotein particles (RNPs) and the post-transcriptional regulation of gene expression. Nearly all of the functions discovered for RNA involve binding of proteins (Haynes, 1999). Several proteins use a β-sheet surface to bind RNA and others insert an α-helix into the widened groove of a non-canonical RNA helix. Distortion or rearrangement of the RNA structure by bound protein is a common feature of their interactions.

In contrast to the DNA interactions, RNAs differ from helical DNA in several ways that affect the possibilities for protein recognition (Draper, 1995) such as:

- Helical segments are interspersed with bulge, internal and hairpin loops that contain noncanonical base pairs and unstacked bases.
- Helices are A-form and typically less than a full turn in length. A-form helices have a very deep and narrow major groove, and it has been supposed that this steric restriction dictates protein recognition in the shallow minor groove. However, the

major groove next to loops or distortions introduced by noncanonical pairs may be quite accessible.

- Flexible tertiary structures may link different parts of an RNA chain to create complex shapes for protein interactions As the consequence, RNA offers a richer diversity of hydrogen bondings, stacking configurations and geometries than DNA for protein recognition.

In addition, mismatches and loops in RNA make them easier for structural deformation upon protein bindings. Consequently, most of the known sequence-specific RNA-binding proteins recognize single-stranded regions and hairpin loops where the functional groups on the bases are accessible. RNA double-stranded regions are recognized only when structural distortions generated by internal loops or bulges allow access to the major groove of the double helix. Furthermore, the diversity of RNA structures favors the recognition of unique shapes and charge distribution of different RNAs by the specific proteins.

The binding specificity between aminoacyl-tRNA synthetases (aRS's) and their cognate tRNAs is one of the most intriguing questions in biochemistry. The fidelity of translation of the genetic code depends on accurate aminoacylation. All tRNAs must have sufficiently similar 3D structures in order to fit interchangeably into the translation apparatus. At the same time, each of the 20 synthetases must be able to select its cognate isoacceptor tRNA from the tRNAs present in a cell, so that only correct tRNA is aminoacylated accurately. tRNA synthetase and tRNA binding involves a large interface area (~3000 Å^2 or ~20% of tRNA accessible surface) with a modest affinity ($K_d \approx 10^{-6}$ M). Differences in binding energy are relatively small for synthetases binding different acceptor tRNAs. The acceptor discrimination occurs mainly at the level of the efficiency of aminoacylation. Thus tRNA-induced conformational changes in the synthetase appear to be a primary mechanism of isoacceptor discrimination. The structural complementarity expressed by a large interface area may provide the free energy necessary for these conformational changes and allow the aRS to differentiate tRNAs from other cellular RNAs (Varani, 1997).

The approach, known as the 'identity swap' experiment, is to determine the minimum number of base changes needed to cause one tRNA species to be selectively aminoacylated by a noncognate synthetase. For the majority of synthetases, the identity elements consist of one or more of three features:

1. at least one base of the anticodon;
2. one or more of the last three bps in the acceptor stem; and
3. the so-called discriminator base (position 73) between the acceptor stem and the CCA terminus, which is always invariant among the isoacceptors of an amino acid.

X-ray crystallographic studies of synthetase complexes (Rould *et al.*, 1991) suggest that these identity elements are involved in the interactions between tRNAs and their cognate synthetases. However, the enzymes may differ in their bindings to the anticodon loop either from the major groove side or the minor groove side. The complexity of the aRS recognition is further demonstrated by the synthetase that uses anticodons from distinct groups and tRNA with an extra variable arm (Biou *et al.*, 1994). In this case, the contacts are made with almost all of the RNA backbone, and the recognition appears to be largely determined by shape, especially that of the variable arm.

10.6 MOLECULAR RECOGNITION IN CARBOHYDRATE-LECTIN INTERACTION

Carbohydrate-protein interactions participate in a wide variety of biological and pathological events. Carbohydrate-binding proteins, excluding enzymes and immunoglobulins, are generally called lectins. Lectins are a useful tool for studying carbohydrate recognition (Lis and Sharon, 1986), e.g. differentiating cell types by cell surface carbohydrates.

10.6.1 Classification and structures of lectins

Lectins are cell-agglutinating and sugar-specific proteins. They are widely distributed and are involved in numerous cellular processes such as host-pathogen intervention, targeting of proteins within cells and cell-cell interactions. Major applications of lectins are:

- cell identification and separation;
- detection, isolation and structural studies of glycoproteins;
- investigation of carbohydrate on cells and subcellular organelles, histochemistry and cytochemistry;
- mapping of neuronal pathways;
- mitogenic stimulation of lymphocytes;
- purging of bone marrow for transplantation; and
- studies of glycoprotein synthesis (Sharon and Lis, 2004).

While lectins are structurally diverse, it is possible to group many of them into distinct families of homologous proteins that share common structural properties (Table 10.6).

Legume lectins are isolated from various leguminous plants, mostly from seeds. They include Man/Glc specific concanavlin A (Con A), pea lectin, *Lathyrus ochrus* lectin (LOL), and Gal preferred *Erythrina corallodendron* lectin (EcorL), soybean allutinin (SBA), representing a family of best studied lectins (Derewenda, *et al.*, 1989; Bourne *et al.*, 1992; Dessen *et al.*, 1995). They generally consist of two or four subunits of 25–30 kDa, each with a single carbohydrate-binding site. The subunits of all these lectins are in the shape of a half dome, with the combining site forming a shallow depression at its apex (Bourne *et al.*, 1992). They exhibit remarkable homologies with over 20% of invariant amino acid residues, including several of those involved in binding of carbohydrates that requires bound Ca^{2+} and/or Mn^{2+}. The overall antiparallel β structure is well conserved within this family of proteins with differences being confined primarily to the loop regions that connect the β strands. Metal ions play dual roles by maintaining the integrity of the subunits of the lectin and positioning amino acid residues for carbohydrate binding, though they do not interact directly with the carbohydrate molecule. The database of lectin structures is accessible at http://www.cermav.cnrs.fr/databank.

The best characterized cereal lectin is wheatgerm agglutinin (WGA) that binds siallyllactose (NeuNACα2→3Galβ1→4Glc). It is a dimer of two identical 18 kDa subunits, each consisting of four homologous domains. These domains are similarly folded, with each having four identically positioned disulfide bridges. Wheatgerm agglutinin is unusual in not having detectable secondary structure (neither α-helix nor β-strand) (Wright, 1990). Another feature is the presence of at least two independent, noncooperative binding sites per subunit. These sites are located at the interface between the subunits that form the molecular dimer of the lectin (i.e. four combining sites per dimer).

TABLE 10.6 Lectins and families

Lectins	Example(s)	Defining features in protein sequence	Carbohydrate recognition	Metal ions
Plant lectins:				
Legume	Con A, SBA		Var: Man/Glc or Gal	Ca^{2+}, Mn^{2+}
Cereal	WGA		GlcNAc, NeuNAc	
Bulb	GNA		αMan	
Animal lectins:				
C-type	Selectins	C-type lectin sequence motif	αMan, αGlcNAc, Fuc, or Gal	Ca^{2+}
S-type	Galectins	S-type lectin sequence motif	βGal	
P-type	M6PRs	Unique repeating motif	Man6P	var
I-type	Sialoadhesin	Immunoglobulin-like domains	NeuNAc	
Pentraxins	SAP	Multimeric binding motif	Var	Yes
Others				
Hy BP	CD44	Sequence homology	Hyaluronan	
Heparin BP	HBPs	Basic amino acid clusters	Heparin and heparin sulfate	
Viral lectins				
Influenza virus	HA		NeuNAc	
Polyoma virus	Viral protein		NeuNAc	
Bacterial lectins				
Toxins	Enterotoxin	Diverse	Gal	

Notes: 1. C-Type lectins include selectins, S-Type and P-type lectins are also known as galectins and mannose-6-phosphate receptors (M6PRs) respectively.

2. Others include carbohydrate binding proteins of animal origins such as hyaluronan BPs (e.g. CD44) which bind hyaluronan and heparin BPs which bind heparin and heparin sulfate.

3. Abbreviations used: BP, binding proteins; ConA, Concanavalin A; GNA, snowdrop lectin; HA, hemagglutinin; HyBP, hyaluronan-binding protein; MBP, mannose-binding protein; SAP, serum amyloid P component; SBA, soybean agglutinin; var, variable; WGA, wheatgerm agglutinin.

Some animal lectins serve as components of the innate host defense system by selectively binding to the surface of potential bacterial and viral pathogens and initiating steps toward their neutralization. Other animal lectins mediate adhesion between animal cells, sorting newly synthesized glycosylated proteins within the luminal compartment of the endoplasmic reticulum, and endocytosis of selected subsets of circulating glycoproteins. Although the number of animal lectins continues to increase, most of the known lectins fall into one of six major groups:

1. C-type or Ca^{2+}-dependent lectins including selectins;
2. S-type of Gal-binding galectins;
3. P-type mannose-6-phosphate receptors;
4. I-type including sialoadhesins and immunoglobulin-like sugar-binding proteins;
5. pentraxins; and
6. other carbohydrate-binding proteins including L-type lectins (related in sequence to legume lectins).

The C-type animal lectins include endocytic receptors such as hepatic lectins, macrophage receptors and selectins that mediate the adhesion of leukocytes to endothelial cells and target the leukocytes to lymphoid tissues and sites of inflammation. Each of these proteins contains one or more carbohydrate-recognition domains (CRDs) combined with domains responsible for the other functions of the molecule (Hughes, 1992). The CRDs are roughly 120 amino acids long, containing 14 invariant amino acids and another 18 that are conserved in character. All C-type lectins require Ca^{2+} for carbohydrate binding and they exhibit two sets of binding specificities, one set of C-type lectins binding derivatives of Man and GlcNAC (Man-specific) and the other binding Gal and GalNAc (Gal-specific).

Three members of the selectin (C-type) family of adhesion proteins are L selectin, which is expressed on various leukocytes, E selectin, which is expressed on endothelium activated by inflammatory mediators and P selectin, which is stored in alpha granules of platelets and is also expressed on endothelium activated by inflammatory stimuli. They share common structural motifs; i.e. calcium dependent (type C) lectin domain, epidermal growth factor-like domain, short consensus repeat or complement binding protein domain and transmembrane anchoring motif with short cytoplasmic tail. All three selectins mediate Ca^{2+} dependent cell adhesion through the recognition of carbohydrates. The carbohydrate ligands for selectins are complex and contain the probable structures of: NeuNAcα2→3(SO_4-6)Galβ1→4(Fuc1→3)GlcNAc and NeuNAcα2→3Galβ1→4 (Fuc1→3)(SO_4-6)GlcNAc. The complex situations of the carbohydrate recognition is the presence of the glycoprotein scaffolds in selectins due to the presence of various oligosaccharide chains on protein backbones that appear to contribute to the binding avidity and specificity of the selectins.

Two properties common to members of the galectin family are shared characteristic amino acid sequences and affinity for β-galactoside sugars. The amino acid identity in the carbohydrate-binding domains, among different galetins from one mammalian species, ranges from about 20 to 40%. The identity of the same galectin from different mammalian species is 80–90%. All mammalian galectins recognize the same structural determinant on lactose and related β-galactosides. Although the major interaction is with the galactose residue in lactose, an interaction with the glucose residue in lactose is also significant. Substitutions on the β-galactose residue differentially affect interaction with specific galectins, presumably reflecting differences in β-strands, which may contribute to an extended binding site. It might also accommodate consecutive *N*-acetyllactosamine residues as found in polylactosaminoglycans, which are particularly good ligands for many galectins.

I-Type lectins are a family of mammalian lectins containing the immunoglobulin(Ig)-like domain. All are integral membrane proteins, preferentially expressed on the plasma membrane, and some have large cytosolic domains with multiple potential phosphorylation sites (both Ser/Thr and Tyr). Sialoadhesin is found in bone marrow and tissue macrophages, mediating NeuNAc dependent adhesion to various lymphohematopoietic cells. It possesses Ig-like domains and shows considerable N-terminal homology with another member, CD22, which is a cell surface phosphoglycoprotein detected on the majority of resting B cells. CD22 appears to facilitate antigen-dependent B cell triggering by association with the B cell antigen receptor with cytoplasmic tyrosine kinase. It can also induce intercellular adhesion, recognizing ligands on activated lymphocytes, monocytes and endothelial cells. CD22 interaction involves recognition of NeuNAcα2→6Galβ1→4GlcNAcβ–, known to occur on the N-linked oligosaccharides of some cell surface glycoproteins. The absolute requirement of the NeuNAcα2→6 linkage for CD22 contrasts with that of the Neuα2→3 linkage for sialoadhesin.

Some pathogens use carbohydrate recognizing lectins as a means of attachment to eukaryotic cell surfaces such as hemagglutinins of influenza and other viruses. These lectins usually recognize NeuNAc, which is the common sugar constituent of the cell surface glycoproteins. These surface binding lectins tend to achieve the required planar array of carbohydrate-binding sites by arranging their oligomeric subunits in cyclic symmetry. Bacterial toxins that use carbohydrates as cellular receptors also display common structural features (Burnette, 1994).

10.6.2 Lectin-carbohydrate recognition: general

Several of the carbohydrate-binding proteins, though often completely unrelated in their functions, have similar structures that are characterized by the presence of two distinct globular domains with a deep cleft between them. Many ligand binding sites occur in the cleft where the domains approach each other (Quiocho, 1986). Thus carbohydrate-binding proteins can be divided into two major groups according to the topological features of the combining site (Rini, 1995). The proteins constituting group I, such as bacterial periplasmic transport proteins and enzymes, have a buried binding site and engulf the ligand fully upon binding. Proteins such as lectins belonging to group II have a shallow binding site, mostly in the form of a depression on the protein surface.

In spite of the enormous diversity of lectins, two aspects of their organization are noteworthy. First, the sugar-binding activity can be ascribed to a limited portion of most lectin molecules, typically a globular carbohydrate-recognition domain (CRD) of less than 200 amino acids. Second, CRDs of many lectins are related in the amino acid sequence to each other, so that most of them can be organized into a relatively small number of classes. From the similar amino acid sequences within each of these groups, it appears that the members of each group also share fundamental three-dimensional structural features. Many of these overall features have been reviewed (Rini, 1995).

Nature seems able to construct saccharide-binding sites on very different frameworks, as exemplified by the legume lectins with their large β structures and low cysteine content on one hand, and WGA (cereal lectin) with its high cysteine content and virtual absence of regular structures on the other. Chemical groups involved directly in binding are diverse. Carbohydrates interact with lectins through hydrogen bonds, metal coordination (metal-dependent lectins), van der Waals and hydrophobic interactions. The contributions of these chemical interactions in carbohydrate-lectin recognition (Weis and Drickamer, 1996) will be considered.

10.6.2.1 Hydrogen bonding. The availability of a large number of hydroxyl groups on carbohydrates renders them obvious partners in complex networks of hydrogen bonds, usually formed by cooperative hydrogen bonds in which the hydroxyl serves both as a donor and an acceptor. Cooperative hydrogen bonding is characteristic of the interaction of lectins with saccharide hydroxyls. The amino acids most commonly involved in hydrogen bonds with carbohydrates are known to be Asp, Asn > Glu > Arg, His, Trp, Lys > Tyr, Gln > Ser, Thr. Generally one acidic side chain (Asp and Glu) from lectins is used as a hydrogen bond acceptor from one or two sugar OH's. Hydrogen-bond donors come primarily from main-chain amide groups and the side-chain amide group of Asn, and less frequently, Gln. Charged side-chain donors (Arg, Lys and His) also participate in hydrogen bonding to some frequency. Protein hydroxyl groups from Tyr, Ser and Thr are less common as either donors or acceptors of hydrogen bonds, because the entropic cost of fixing the rotamers of both glycose and protein hydroxyls (to geometry of planar donors and acceptors in hydrogen bonds) is likely expensive.

In some cases, a pair of vicinal glycose OH's or one OH and the ring oxygen interact with two functional groups in a single amino acid side chain or with consecutive main-chain amide groups. Since spacing (~2.8 Å) of vicinal hexopyranose OH's in either equatorial/equatorial or equatorial/axial configuration is within the range of hydrogen bonding, excellent hydrogen-bond geometry can be achieved between planar amino acid side chains and vicinal glycose OH's in this configuration. Vicinal axial/axial OH's in hexopyranoses are separated by about 3.7 Å (Vyas, 1991) and do not make multiple hydrogen bonds with a single amino acid side chain.

The cyclic oxygen atom of glycoses can act as hydrogen-bond acceptors but cannot participate in cooperative hydrogen bonds. In direct protein interactions, this oxygen usually shares a hydrogen-bond donating amino acid side chain with one of the glycose OH's. Moreover this oxygen is common to all glycoses and thus cannot be used to distinguish among them. The acetamido moiety of GlcNAc, GalNAc and NeuNAc is often a dominant or significant recognition determinant. Unlike hydroxyls, the amide group and carbonyl oxygen of the acetamido substituent have fixed, planar geometry. The amide group acts as a hydrogen bond donor to planar carbonyl or carboxylate oxygen, while the acetamido oxygen often accepts hydrogen bond from Ser of lectins.

The glycose functionalities that form hydrogen bonds with the lectin are those required for specific recognition and discrimination, whereas those positions that are not used as recognition elements tend to be exposed to solvent and form no direct contacts with the lectin. For example, the specificity for Gal in the galectins is achieved using the axial 4-OH and 6-OH; and equatorial 4-OH would be sterically excluded by Trp68 and would be unable to form the same hydrogen-bond interactions with Arg and His (Liao *et al.*, 1994).

In general, the hydrogen bonds between lectins and essential recognition determinants on carbohydrates are shielded form bulk solvent, Moreover, the use of hydrogen bond donors and acceptors with fixed geometry may be important to specificity. A freely rotating group like the OH of Ser/Thr has some plasticity in the formation of hydrogen bonds with the saccharide and may not be capable of discriminating absolutely between epimeric OH's, therefore are infrequently used in the discriminatory carbohydrate-lectin recognition.

10.6.2.2 Nonpolar interaction. Carbohydrates are highly polar and solvated molecules owing to the presence of the OH's and the cyclic oxygen. Nonetheless, glycoses have significant nonpolar patches formed by the aliphatic protons and carbons at the various epimeric centers, which extend out to the exocyclic 6 position of hexopyranoses as well as the glycerol moiety of neuraminic acids. This patch is observed to pack against the face of one or more aromatic side chains of the protein. In all lectin-Gal complex structures, the apolar patch of the Gal B face (the B face is defined as that side of the ring that gives clockwise numbering along the ring, cf. Rose *et al.*, 1980) packed against the face of Trp or Phe. This arrangement is frequently described as stacking, which implies that the ring of Gal is parallel to the plane of the aromatic ring. Superposition of the bound Gal in various structures reveals that the aromatic rings form an angle between 17° and 52° with the least-square plane through the pyranose ring and exocyclic carbon, with an average angle of 32° (Kolatkar and Weis, 1996). In addition, the methyl group of the acetamido moiety of GlcNAc, GalNAc and NeuNAc often interacts with an aromatic ring in lectins that specifically recognize this group. The carbon backbone of the glycerol moiety of NeuNAc also contributes apolar surface to the protein. The interaction with the delocalized electron cloud of the aromatic ring is energetically significant beyond simply providing a geometrically complementary apolar surface. Most likely, the interaction is driven by the proximity of the aliphatic protons of the sugar ring, which carries a net positive partial charge, and the π-electron cloud of the aromatic ring. Thus the amino acids with aromatic or aliphatic side chains such as Trp, Phe, Tyr, Leu, Val and Ala are usually observed in van der Waals interactions with carbohydrates.

10.6.2.3 Metal ion. Several classes of lectins require divalent metal ions for function. The legume lectins use Ca^{2+} and Mn^{2+} to stabilize the binding site and fix the positions of amino acids that interact with sugar ligands. The Ca^{2+} forms coordination bonds with the side-chain carbonyl oxygen of a conserved Asn while the side chain NH_2 of this Asn donates a hydrogen bond to the carbohydrate ligand. The Ca^{2+} fixes side chains for optimal binding to glycose and stabilizes the general architecture of the binding site. The Mn^{2+} does not coordinate any residues that interact directly with the protein, but instead fixes the Ca^{2+} position. However, the Ca^{2+} of the C-type lectins forms direct coordination bonds with the carbohydrate ligand. One lone pair of electrons from each of the two vicinal OH's forms a coordination bond with Ca^{2+}.

10.6.2.4 Extended site or subsite multivalency and subunit multivalency. Some lectins are specific for only a single monosaccharide unit and do not discriminate on the basis of what other sugars are linked to this recognition element. In other cases, substantial affinity enhancements are observed when additional glycose units are linked to the primary determinant. The binding site that interacts specifically with more than one glycose residue to provide enhanced affinity is termed extended site or subsite multivalency. For example, pea lectin displays higher affinity for oligo-GlcNAc containing α(1,6)-linked Fuc, whereas ConA does not. Other lectins show no significant affinity for monosaccharides, but instead bind specifically to larger oligosaccharides. For example, galectins display marginal affinity for Gal but bind lactose (Galβ1→4Glc) and *N*-acetyllactosamine (Galβ1→4GlcNAc) with the preference for the latter. This specificity appears to be due to the formation of hydrogen bonds by the 3-OH of Glc or GlcNAc with amino acids that are also ligands for the Gal residue. Tighter binding of *N*-acetyllactosamine compared to lactose has been attributed to van der Waals contacts between the acetamido group of GlcNAc and the protein that is absent in the complex with lactose. Water may also mediate the extended site interactions, presumably by forming hydrogen bonds between OH's and ring oxygen to stabilize the oligosaccharide conformation, as well as by mediating interactions between carbohydrate and protein. As carbohydrates display a great deal of conformational heterogeneity, owing to the relatively small energetic barriers to rotation about glycosidic linkages, the ability of water molecules to mediate contacts with the protein surface is undoubtedly important in selecting a particular oligosaccharide conformation from the ensemble that exists in solution. However, steric exclusion of certain oligosaccharide conformations by structural features of the protein surface may be equally important to the overall specificity and function of the lectin.

Lectins can have two clearly separated and thermodynamically independent binding sites in a single protomer. The most well characterized examples are WGA (Wright, 1984) and snowdrop lectin (Hester *et al.*, 1995). The 3D structures and the amino acid sequences reveal internal duplications so that the primary and secondary binding sites are homologous. The multiple independent binding sites of the lectin can give rise to increased affinity for multivalent ligands. There is no evidence of molecular cooperativity for either secondary or primary sites. In certain lectins, such as influenza hemagglutinin and mannose binding protein (MBP), secondary sites having much lower affinity with no sequence or structural homology to the primary site have been identified.

Higher binding affinity is achieved by extending binding sites through additional direct and water-mediated contacts between oligosaccharides and the protein surface. Dramatically increased affinity for oligosaccharide results from clustering of simple binding sites in oligomers of the lectin polypeptides. The geometry of such oligomers in subunit multivalency helps to establish the ability of the lectins to distinguish surface arrays of polysaccharides in some instances and to cross-link glycoconjugates in others. Multivalency is exhibited when several subunits of the same lectin bind to different extensions of a branched carbohydrate, as in the case of the asialoglycoprotein receptor (ASGPR), or to separate carbohydrate chains as in the case of the trimeric MBP and the pentameric cholera toxin. Multiple binding sites recognizing multivalent ligands create an apparent higher affinity called avidity.

10.6.2.5 Conformational change. Lectins undergo few, if any, changes in conformation upon binding to saccharides. However, no global changes in protein structure have been observed. Small chain movements are restricted to the immediate vicinity of the bounded glycose. The glycose binding reduces the amount of conformational variability of certain residues by selecting from the ensemble in solution the conformation that optimizes the contacts. The binding sites in lectins are preformed, with ordered water molecules forming hydrogen bonds with the unliganded proteins in a pattern that closely mimics the hydrogen bonding by glycose hydroxyls (Rini *et al.*, 1993). This effect results from the use of donors and acceptors of the planar hydrogen bonds, which provide fixed geometry for interactions among amino acid residues in the binding site that maintain the conformation of the side chains in the absence of saccharide ligands.

10.6.3 Lectin-carbohydrate recognition: ligand discrimination

The differential binding of glycoses to lectins depends on subtle changes in the disposition of amino acid side chains near the carbohydrate-binding sites. The sites are broad and shallow, with selectivity resulting from three factors:

1) The epimeric OH that distinguishes particular glycoses, such as the 4-OH in Man-type versus Gal-type ligands, is involved in hydrogen bond and sometimes coordination bond to the lectins, so that the binding site directly reads out the differences among these glycoses.
2) A second factor in establishing selectivity is steric exclusion of the disfavored ligands, rather than just loss of certain contacts compared to lectins that bind the same ligands tightly. Conversely, accommodation of multiple ligands by one particular binding site is achieved by leaving exposed those portions of the ligands that would allow them to be distinguished. The formation of alternative, compensating contacts with two different preferred ligands is not observed.
3) Lectins can be insensitive to how a bound glycose is linked in a simple glycoside or an oligosaccharide as long as the attached substituents project away from the lectin surface.

Comparison of lectins with homologous amino acid sequences that bind different saccharide ligands may provide useful information about the interaction mechanism for lectins of family members to discriminate among the different ligands. Two aspects of carbohydrate recognition by lectins will be considered.

10.6.3.1 Discrimination between Gal versus Man/Glc. While the ultimate selectivity of lectins is probably attained by the subsites and subunit multivalency, a major discriminatory factor in carbohydrate recognition is the primary monosaccharide specificity. In general, selectivity are achieved through a combination of hydrogen bonding to the saccharide hydroxyl groups with van der Waal packing, often including packing of a hydrophobic glycose face against aromatic amino acid side chains. Although the key interactions responsible for carbohydrate recognition are common, each family has evolved a unique stereochemistry at the principal combining site in order to discriminate between ligands. The common view is that the selectivity toward a particular target is augmented through multiple binding, by mechanism of additional binding in subsite (or extended site) and/or subunit multivalency (Rini, 1995; Weis and Drickamer, 1996). In subsite binding, one monosaccharide, usually the terminal one, is bound at the primary binding site of the lectin, with additional monosaccharides along the carbohydrate chain bound to secondary subsites on the lectin. This kind of selectivity enhancement is demonstrated in the binding of carbohydrate ligands to legume lectins. Among these plant lectins, concanavalin A (ConA), pea lectins and *Lathyrus ochrus* lectin (LOL) bind primarily to Man/Glc saccharides, whereas *Erythrina corallodendron* lectin (EcorL) and soybean agglutinin (SBA) bind Gal saccharides preferentially. Likewise the C-type animal lectins, MBP and human ASGPR form complexes with complex carbohydrates terminated with Man and Gal respectively. Gal and Glc/Man are C-4 epimers with Gal having an axial 4-OH, whereas Glc/Man having an equatorial 4-OH. Among C-type lectins, the discrimination resulting from the three effects:

1. *Amide positions at the binding site*: Exchanging the position of a single amide and a carboxylate oxygen between two of the four amino acid side chains (Asp, Asn, Glu and/or Gln) that bond both Ca^{2+} and the saccharide ligand is sufficient to invert the saccharide binding selectivity of the binding site.
2. *Packing interactions unique to galactose*: The packing of an aromatic ring against the apolar B face of Gal in unique to the Gal-binding site and no such interaction occurs with Man, although the CRD for the Gal- and Man-specific lectins are similar. Such stacking of aromatic residues is a feature of lectin-Gal interactions that contributes to the stability and specificity of the complexes formed.
3. *Steric exclusion*: Exclusion plays an important role in determining selectivity between Man and Gal.

Two conserved residues, Asp and Asn that coordinate Ca^{2+} to the protein, also form hydrogen bonds with the saccharide in the combining site of legume lectins. The two residues have an identical spatial disposition in these lectins. However, the saccharide is rotated so that a different set of OH's is involved in their interactions, i.e. the 4- and 3-OH's of Gal versus the 4- and 6-OH's of Man.

The donors and acceptors of the axial and equatorial 4-OH tend to cluster in two segregated regions of their respective domes, formed by the spanning glycose hydroxyls (Elgavish and Shaanan, 1997). The donors of the axial 4-OH and the acceptors of the equatorial 4-OH span the segment between *gauche*$^-$ and *trans* conformations around the 4-OH—C-4 bond, whereas the axial 4-OH acceptors and equatorial 4-OH donors span the *trans* and *gauche*$^+$ segment. The unique clustering mode of donors and acceptors around the 4-OH as opposed to the even distributions around other hydroxyls, emphasizes the importance of the stereochemistry around the sugar C-4 position for Gal versus Glc/Man specificity. The disposition of the aromatic residues interacting with the carbohydrate is also coupled with that of the donors. The overlapping aromatic groups of Gal-specific proteins and the respective donors face opposite sides of the hexose ring. The disposition of the aromatic groups with respect to the ligand is optimal in the Gal-binding proteins. The requirements for the disposition of donors/acceptors necessary to recognize the axial 4-OH of Gal are clearly distinguishable from the requirements for binding the equatorial

4-OH of Glc/Man. It appears that a selection to favor the correct carbohydrate in the combining site is achieved by matching the C-4 epimer with the given constellation of 4-OH donors/acceptors. When a match is established, the binding energy is enhanced by contributions from other ligand- and protein-dependent factors, such as additional hydrogen bonds, van der Waals interactions and aromatic stacking. While satisfying these basic epimeric dependent hydrogen-bonding constraints around C-4 may be crucial for the initial selection of primary ligand, additional ligand-dependent hydrogen bonds, van der Waals and hydrophobic interactions are obviously essential for highly specific binding to take place.

10.6.3.2 Interaction with NeuNAc: an example. The sialic acid (NeuNAc) of sialyllactose is bound to wheatgerm agglutinin (WGA) by several polar interactions and a large number of nonpolar contacts with six amino acid side chains forming two different subunits. All the NeuNAc ring substituents are involved, i.e. the acetamido and carboxyl groups as well as the hydroxyl groups attached to the pyranose ring and the glycerol side chain. The acetamido group makes the largest number of contacts. Its carbonyl is hydrogen bonded to the hydroxyl of Ser62 and its amide to the carbonyl of Glu115. In addition, it forms five to seven van der Waals contacts with the phenyl ring of Tyr73; the latter side chain provides a nonpolar contact for the methyl group of the acetamido residue of the sialic acid. The adjacent ring hydroxyl (4-OH) is fixed by hydrogen bonds with the hydroxyl of the same Tyr. The carboxylate group of the *N*-acetylneuraminic acid is within hydrogen-bonding distance of the hydroxyl of Ser114. Numerous van der Waals contacts stabilize the orientation of the sugar ring through nonpolar stacking interactions with the aromatic side chain of Tyr66. A third aromatic side chain, that of Tyr64, interacts through nonpolar contacts with the glycerol tail of the sialic acid. Although all the ring substituents of the NeuNAc are in contact with protein side chains, it is the *N*-acetyl group and the adjacent 4-OH that are the essential specificity determinants for the two monosaccharides, *N*-acetylglucosamine and *N*-acetylneuraminic acid, that can wind to WGA. They provide a cluster of three spatially close hydrogen bonds and a hydrophobic contact (acetamido-CH_3 with the aromatic ring of Tyr73) in the least exposed part of the binding cavity, where the conformation of protein is most stable.

10.7 REFERENCES

ALZARI, P.M., LASCOMBE, M.-B. and POLJAK, R.J. (1988) *Annual Reviews in Immunology*, **6**, 555–80.

AMIT, A.G. *et al.* (1986) *Science*, **233**, 747–53

AMZEL, L.M. and POLJAK, R.J. (1979) *Annual Reviews in Biochemistry*, **48**, 961–97.

BIOU, V., YAREMCHUK, A., TUKALO, M. and CUSACK, S. (1994) *Nature*, **263**, 1404–10.

BOURNE, Y., ROUGÉ, P. and CAMBILLAU, C. (1992) *Journal of Biology and Chemistry*, **267**, 197–203.

BRENNAN, R.G. and MATTEWS, B.W. (1989) *Trends in Biochemical Science*, **14**, 286–90.

BURTON, D.R. (1990) *Trends in Biochemical Sciences*, **15**, 64–9.

DAVIES, D.R. and METZGER, H. (1983) *Annual Reviews in Immunology*, **1**, 87–117.

DAVIES, D.R., PADLAN, E.A. and SHERIFF, S. (1990) *Annual Reviews in Biochemistry*, **59**, 439–73.

DEREWENDA, Z., YARIV, J., HALLIWELL, J.R. *et al.* (1989) *EMBO Journal*, **8**, 2189–93.

DESSEN, A., GUPTA, D., SABESAN, S. *et al.* (1995) *Biochemistry*, **34**, 4933–42.

DRAPER, D.E. (1995) *Annual Reviews in Biochemistry*, **64**, 593–620.

ELGAVISH, S. and SHAANAN, B. (1997) *Trends in Biochemical Science*, **22**, 462–7.

FAIRALL, L., RHODES, D. and KLUG, A. (1986) *Journal of Molecular Biology*, **192**, 577–91.

GOLEMIS, E. (ed.) (2002) *Protein-Protein Interactions*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor.

HARDING, S.E. and CHOWDHRY, B.Z. (eds) (2001) *Protein-Ligand Interactions: Structure and Spectroscopy*, Oxford University Press, Oxford.

HARRISON, S.C. and AGGARWAL, A.K. (1990) *Annual Reviews in Biochemistry*, **59**, 933–69.

HAYNES, S.R. (1999) *RNA-Protein Interaction Protocols*, Humana Press, Totowa, NJ.

HESTER, G., KAKY, H., GOLDSTEIN, I.J. and WRIGHT, C.S. (1995) *Nature Structural Biology*, **2**, 472–9.

HORN, A. and BÖRNIG, H. (1969) *FEBS Letters*, **3**, 325–9.

HUGHES, R.C. (1992) *Current Opinions in Structural Biology*, **2**, 687–92.

JANIN, J., MILLER, S. and CHOTHIA, C. (1988) *Journal of Molecular Biology*, **204**, 155–64.

JOHNSON, P.F. and MCKNIGHT, S.L. (1989) *Annual Reviews in Biochemistry*, **58**, 799–839.

KEHOE, J.E. and CAPRA, J.D. (1971) *Proceedings of the National Academy Sciences, USA*, **68**, 2019–21.

KIM, J.L., NIKOLOV, D.B. and BURLEY, S.K. (1993) *Nature*, **365**, 520–7.

KLEANTHOUS, C. (ed.) (2000) *Protein-Protein Recognition*, Oxford University Press, Oxford.

KLOTZ, I.M. and HUNSTON, D.L. (1971) *Biochemistry*, **16**, 3065.

KOLATKAR, A. and Weis, W.I. (1996) *Journal of Biology and Chemistry*, **271**, 6679–85.

KOSHLand, D.E., NÉMETHY, G. and FILMER, D. (1966) *Biochemistry*, **5**, 365–85.

LandSCHULTZ, W.H., Johnson, P.F. and McKnight, S.L. (1988) *Science*, **240**, 1759–64.

LIAO, D.I., KAPADIA, G., AHMED, H., VASTA, G.R. and Herzberg, O. (1994) *Proceedings of the National Academy Sciences, USA*, **91**, 1428–32.

LILLEY, D.M.J. (ed.) (1995) *DNA-Proteins: Structural Interactions*, Oxford University Press, Oxford, UK.

LIS, H. and SHARON, N. (1986) *Annual Reviews in Biochemistry*, **55**, 35–67.

MONAD, J., WYMAN, J. and CHANGEUX, J.P. (1965) *Journal of Molecular Biology*, **12**, 88–118.

PABO, C.O. and SAUER, R.T. (1984) *Annual Reviews in Biochemistry*, **53**, 293–321.

PADLAN, E.A. *et al.* (1989) *Proceedings of the National Academy Sciences, USA*, **86**, 5938–42

PHILLIPS, S.E. (1991) *Current Opinions in Structural Biology*, **1**, 89–98.

QUIOCHO, F.A. (1986) *Annual Reviews in Biochemistry*, **55**, 287–315.

RICE, P.A., YANG, S.-W., MIZOGUCHI, K. and NASH, H.A. (1996) *Cell*, **87**, 1295–306.

RICHARD, J. and CORNISH-BOWDEN, A. (1987) *European Journal of Biochemistry*, **166**, 255–72.

RINI, J.M. (1995) *Annual Reviews in Biophysical Biomolecular Structure*, **24**, 551–7.

RINI, J.M., HARDMAN, K.D., EINSPAHR, H. *et al.* (1993) *Journal of Biology and Chemistry*, **268**, 10126–32.

ROSE, I.A., HANSON, K.R., WILKINSON, K.D. and WIMMER, M.J. (1980) *Proceedings of the National Academy Sciences, USA*, **77**, 2439–41.

ROULD, M.A., PERONA, J.J. and STEITZ, T.A. (1991) *Nature*, **352**, 213–8.

SAENGER, W. and HEINEMANN, U. (ed.) (1989) *Protein-Nucleic Acid Interaction*, CRC Press, Boca Raton, Fl.

SCHLEIF, R. (1988) *Science*, **241**: 1182–7.

SCHWABE, J.W.R. and RHODES, D. (1991) *Trends in Biochemical Sc*ience, **16**, 291–6.

SEEMAN, N.C., ROSENBERG, J.M. and RICH, A. (1976) *Proceedings of the National Academy Sciences, USA*, **73**, 804–8.

SHAKKED, Z., GUZIKEVICHGUERESTEIN, G., FROLOW, F. *et al.* (1994) *Nature*, **368**, 469–73.

SHARON, N. and LIS, H. (2004) *Glycobiology*, **14**, 53R–62R.

SHERIFF, S. *et al.* (1987) *Proceedings of the National Academy Sciences, USA*, **84**, 8075–9

STEINHARDT, J. and REYNOLDS, J.A. (1969) *Multiple Equilibria in Proteins*, Academic Press, New York.

STEITZ, T.A. (1993) *Structural Studies of Protein-Nucleic Acid Interaction: The Sources of Sequence-Specific Binding*, Cambridge University Press, Cambridge, UK.

STEITZ, T.A., OHLENDORG, D.H., MCKAY, D.B. *et al.* (1982) *Proceedings of the National Academy Sciences, USA*, **79**, 3097–100.

SUCK, D., LAHM, A. and OEFNER, C. (1988) *Nature*, **332**, 464–8.

TRAVERS, A. (1993) *DNA-Protein Interactions*, Chapman & Hall, London.

TULIP, W.R. *et al.* (1989) *Cold Spring Harbor Symposium, Quantum Biology*, **54**, 257–63

VAN OSS, C.J. and VAN REGENMORTEL, M.H.V. (eds) (1994) *Immunochemistry*. Marcel Dekker, New York.

VARANI, G. (1997) *Account Chemical Research*, **30**, 189–95.

VYAS, N. (1991) *Cun. Opin. Stuct. Biol.* **1**, 732–40.

WARREN, R.W. and KIM, S.H. (1978) *Nature*, **271**, 130–5.

WEBER, G. (1992) *Protein Interactions*, Chapman & Hall, London.

WEIS, W.I. and DRICKAMER, K. (1996) *Annual Reviews in Biochemistry*, **65**, 441–73.

WEISS, J.N. (1997) *FASEB Journal*, **11**, 835–41.

WRIGHT, C.S. (1984) *Journal of Molecular Biology*, **178**, 91–104.

WRIGHT, C.S. (1990) *Journal of Molecular Biology*, **215**, 635–51.

WU, T.T. and KABAT, E.A. (1970) *Journal of Experimental Medicine*, **132**, 211–50.

World wide webs cited

BIND:	http://www.bind.ca
DIP:	http://dip.doe-mbi.ucla.edu
Epitome:	http://www.rostlab.org/services/epitome
IMGT:	http://cines.fr
IntAct project (EBI):	http://ebi.ac.uk/intact
KABAT:	http://immuno.bme.nwu.edu/
Lectin structure database:	http://www.cermav.cnrs.fr/databank
ProNIT:	http://gibk26.bse.kyutech.ac.jp/jouhou/pronit/pronit.html

CHAPTER 11

BIOMACROMOLECULAR CATALYSIS

11.1 BIOCATALYST: DEFINITION AND CLASSIFICATION

Catalysis refers to a facilitated chemical transformation under the influence of a catalyst, while the catalyst itself is not chemically modified and appears in the stoichiometric equation as both reactant (A) and product (P). Thus the catalysis is intimately associated with reaction velocity and stoichiometry of the reaction. For a simple uncatalyzed reaction,

$$\mathrm{A} \rightarrow \mathrm{P}$$

the rate (v_u) expression is

$$v_u = -d[\mathrm{A}]/dt = k_u[\mathrm{A}]$$

where k_u is the rate constant characteristic of the reaction.

In the presence of a catalyst (C):

$$\mathrm{A} + \mathrm{C} \rightarrow \mathrm{P} + \mathrm{C}$$

the rate (v_c) of the catalyzed reaction becomes

$$v_c = -d[\mathrm{A}]/dt = k_c[\mathrm{A}][\mathrm{C}] = k_{cat}[\mathrm{A}]$$

Since the catalyst is not used up ([C] = constant), its concentration may be combined with the rate constant, k_c to give k_{cat}, which is the apparent rate constant of the catalyzed reaction. Thus a catalyst is a substance that alters the velocity of the chemical reaction without appearing in the end product of the reaction (Glasstone *et al.*, 1941). To express catalysis in mechanistically meaningful and useful terms, the role of a catalyst is to make alternate reaction pathway(s) available to the system. The requirements for this new pathway are:

- It circumvents the rate-limiting step of the original (uncatalyzed) pathway: The catalyzed reaction is faster than the uncatalyzed one, thereby satisfying the definition of a catalyst as an accelerator.
- It leads to products that differ from the original set only by the presence of catalyst: The effect of the catalyst is limited to the velocity. It does not affect the course of the reaction, i.e. the stoichiometry.
- The catalyst is easily removed from the products by exchange with reactants. The catalyst is regenerated and therefore is effective when present in small quantities relative to the reactants.

The effectiveness of a catalyst is assessed by the relative velocities of the original reaction and the alternate pathway(s). Each pathway usually has some particular step that controls the overall rate. The rate of this step is determined by its free energy of activa-

---

Biomacromolecules, by C. Stan Tsai
Copyright © 2007 John Wiley & Sons, Inc.

tion, which is the highest free energy barrier along the route from reactants to products. The most important factor in catalytic efficiency is the amount by which these free energies of activation differ for the catalyzed and uncatalyzed reactions. Thus catalysis of a chemical reaction is acceleration by a substance known as a catalyst that is not consumed in the overall reaction. The catalyst usually functions by interacting with the reactant(s) to yield a set of species that can react by an alternate pathway involving a lower free energy of activation to give the product(s) and to regenerate the catalyst. For equilibrium reactions, it is important to note that a catalyst speeds up the attainment of equilibrium but does not change the position of equilibrium. Thus the equilibrium constant is not changed in going from the uncatalyzed reaction to the catalyzed reaction, as the catalyst only speeds up on approaching to the state of equilibrium.

Most of the chemical reactions that take place in biological systems are accelerated above the basal chemical reaction rate. These accelerated reactions are promoted by biocatalysts (Bommarius and Riebel, 2004; Griengl, 2000), which can be defined as macromolecules that:

a) accelerate the rate of a reaction by providing an alternative reaction pathway with a lower activation energy;

b) are highly specific with respect to the substrates they act upon and the product they generate; and

c) are not consumed in the reaction.

Other biochemical reactions are quasi-catalytic; they involve biomacromolecules that meet the criteria (a) and (b) but not (c), i.e. only one turnover. For example, in the case of self-splicing, the intramolecular catalysis that occurs at a specific site within a biomacromolecule is accelerated by the folded structure of the molecule but the macromolecule is physically changed in the reaction. In this text, biocatalysts or catalytic biomacromolecules refer to:

- *Protein*s: Enzymes (natural catalytic proteins), abzymes (engineered catalytic antibodies) and self-splicing proteins.
- *Nucleic acids*: Ribozymes (catalytic RNA) and deoxyribozymes (catalytic DNA)
- *Glycan*s: Glycozymes (catalytic glycans)?

The vast majority of biochemical reactions are catalyzed by enzymes (Bagg, 2004; Copeland, 2000), which are proteins fulfilling all the three criteria of being biocatalysts; efficiency (a), specificity (b) and multiple turnover (c). The term enzymes is reserved for the natural catalytic proteins in this text. Although few quasi-catalytic enzymes are known such as type I restriction endonuclease, poly(ADP-ribose) synthetase and transmethylase for O^6-methylguanine mediate specific reactions with large acceleration but are inactivated in the reaction.

Enzymes are usually named with reference to the substrate/reaction they catalyze. It is customary to add the suffix *–ase* to the catalyzed reaction and the name of its major substrate. The Enzyme Commission (EC) has recommended nomenclature of enzymes based on the six major types of enzyme-catalyzed reactions:

EC 1: *Oxidoreductases* catalyze oxidation-reduction reactions: The substrate that is oxidized is regarded as hydrogen donor. The systematic name is based on donor:acceptor oxidoreductase. The name *dehydrogenase* is recommended wherever possible and *reductase* is used as an alternative. *Oxidase* is used only when O_2 is the acceptor.

EC 2: *Transferases* catalyze group transfer reactions: The enzymes catalyze transfer of a group from one compound (donor) to another compound (acceptor). The systematic name is formed according to the scheme donor:acceptor grouptransferase. *Acceptor grouptransferase* or *donor grouptransferase* is the recommended name.

EC 3: *Hydrolases* catalyze hydrolytic reactions: These enzymes catalyze the hydrolytic cleavage of C—O, C—N, C—C and other bonds (e.g. phosphoric anhydride bonds), although the systematic name always includes hydrolase. The name of the substrate suffixed with *–ase* is used in many cases (especially common names). A number of hydrolases are known to catalyze not only hydrolytic removal of a particular group from their substrate but likewise transfer of this group to suitable acceptor molecules. However, in most cases the reaction with water as the acceptor was discovered earlier and are now considered as hydrolases.

EC 4: *Lyases* catalyze cleavage and elimination reactions: The enzymes cleave C—C, C—O, C—N and other bonds by elimination, leaving double bonds, or conversely adding groups to double bonds. The systematic name is formed according to the pattern, substrate group-lyase (the hyphen is an important part of the name and should not be omitted). Some names such as *decarboxylase* (elimination of CO_2), *dehydratase* (elimination of water) are used. In some cases, the reverse reaction is more important, or the only one demonstrated, *synthase* (not synthetase) might be used.

EC 5: *Isomerases* catalyze isomerization reactions: These enzymes catalyze geometric or structural changes within one molecule. According to the type of isomerism, they may be called *racemases*, *epimerases*, *cis-trans-isomerases*, *isomerases*, *tautomerases*, *mutases*, and so on.

EC 6: *Ligases* catalyze synthetic reactions: These enzymes catalyze the joining together of two molecules commonly coupled with the hydrolysis of a pyrophosphate bond in nucleotide triphosphates. The systematic names are formed on the system, X : Y ligase. Use of the term, *synthetase* is recommended.

Thus, the EC numbers provide unique identifiers for enzyme functions, and give us useful keyword entries in database searches (http://www.chem.qmw.ac.uk/iubmb/enzyme/). Enzyme nomenclature/common names and properties are also available at ENZYME (http://www.expasy.org/enzyme) and BRENDA (http://www.brenda.uni-koeln.de). IntEnz (http://www.ebi.ac.uk/intenzy) is the integrated enzyme database and enzyme nomenclature. Table 11.1 lists some enzyme resource sites providing general information.

TABLE 11.1 General enzyme resource sites

Web site	URL
IUBMB: Enzyme classification	http://www.chem.qmw.ac.uk/iubmb/enzyme/
IntEnz: Classification, nomenclature	http://www.ebi.ac.uk/intenz/index.html
ENZYME DB: General information	http://www.expasy.ch/enzyme/
Enzyme Structure database: Structures	http://www.biochem.ucl.ac.uk/bsm/enzyme/index.html
PROCAT: Enzyme active site	http://www.biochem.ucl.ac.uk/bsm/PROCAT/PROCAT.html
LIGAND: Enzyme reactions	http://www.gebine.ad.jp/dbget/ligand.html
EzCatDB: Catalytic mechanism database	http://mbs.cbrc.jp/EzCatDB/
Brenda: Enzyme properties	http://www.brenda.uni-koeln.de/
EMP: General, literature summary	http://wit.mcs.anl.gov/EMP/
TECRdb: Reaction thermodynamics	http://xpdb.nist.gov/enzyme_thermodynamics/
KEGG: Enzyme mediated pathways	http://www.kegg.com/kegg/pathway/map/
PathDB: Primary metabolic pathways	http://www.ncgr.org/software/pathdb/
UM-BBD: Secondary/xenometabolism	http://umbbd.ahc.umn.edu/index.html
Promise: Prosthetic group/metal enzymes	http://bmbsgi11.leads.ac.uk/promise/

Immunoglobulins possess high affinity and unmatched structural specificity toward virtually any molecules. Appropriate antigens have been used to elicit monoclonal antibodies called catalytic antibodies or abzymes that catalyze a variety of chemical reactions (Lerner *et al.*, 1991) including hydrolysis, redox reaction, cyclization, rearrangement and elimination reactions. Self-splicing protein is autocatalytic, mediating the post-translational reaction that involves a precise excision of an internal segment (intein) from a protein precursor and a concomitant ligation of the flanking regions (exteins) resulting in the production of two protein fragments (Cooper and Stevens, 1995).

Some nucleic acids are capable of self-splicing. These catalytic DNA and RNA are known as deoxyribozymes (Li and Breaker, 1999; Sheppard *et al.*, 2000) and ribozymes (Doherty and Doudna, 2000; Scott and Klug, 1996) respectively. The 3′- and/or 2′-hydroxyls of DNA/RNA serve as a catalytic site that invariably requires a metal ion for the catalytic activity. Deoxyribozymes are quasi-catalytic while ribozymes can be catalytic, e.g. ribonuclease P (RNase P) as well as quasi-catalytic, e.g. introns and hammerhead RNAs. RNase P resources are maintained at http://www.mbio.ncsu.edu/RnaseP/home.html

Glycans possess hydroxyl and acetamido groups, which are potentially capable of self-splicing. However, the weak nucleophilicity of these groups and the rigidity of the glycan chains may greatly negate self-splicing plausibility of glycozymes, though neighboring group assistance in the glycosidic cleavage/transfer by the 2′-acetamido and hydroxyl groups has been reported (Piszkiewicz and Bruice, 1967).

11.2 CHARACTERISTICS OF ENZYMES

11.2.1 Enzymes: Catalytic proteins

Enzymes are globular proteins whose sole function is to catalyze biochemical reactions. The most important properties of all enzymes are their catalytic power, specificity and capacity to regulation. The characteristics of enzymes (Bagg, 2004; Copeland, 2000; Kuby, 1991; Price and Stevens, 2000) can be summarized as:

- *Enzymes are proteins*: The term, enzymes refer to natural biological catalysts, which are proteins with molecular weights generally ranging from 1.5×10^4 to 10^8 Da.
- *Cofactor requirement*: Some enzymes require cofactors for the activities. These enzymes that require covalent cofactors (prosthetic groups e.g. heme in cytochromes) or noncovalent cofactors (coenzymes e.g. $NAD(P)^+$ in dehydrogenases) for activities, are called haloenzymes (or simply enzymes). The protein molecule of a haloenzyme is termed proenzyme. The prosthetic group/coenzyme dictates the reaction type catalyzed by the enzyme and the proenzyme determines the substrate specificity.
- *Remarkable properties of enzymes*: Enzymes display a number of remarkable properties in comparison with chemical catalysts. The three most outstanding ones are their high catalytic efficiency (catalytic power), unmatched specificity (discriminatory power) and the extent to which their catalytic activity can be regulated (controllability).
- *Catalytic efficiency* (*Enzymes increase the rate but do not influence the equilibrium of biochemical reactions*): Enzymes are highly efficient in their catalytic power, displaying rate enhancement of 10^{10} to 10^{15} times those of uncatalyzed reactions (Radzicka and Wolfenden, 1995) without changing the equilibrium constants of the reactions.

- *Enzyme specificity* (*Enzymes exhibit a high degree of specificity for their substrates and reactions*): Enzymes are highly specific, both in the nature of the substrate(s) that they utilize and in the types of reactions that they catalyze.
- *Enzyme active site*: The binding and catalytic actions of an enzyme take place in a small region(s) of protein molecules known as binding site and catalytic site respectively. The active site of an enzyme includes both the binding and catalytic sites (Koshland, 1958; Vallee and Riordan, 1969). It is a three-dimensional entity generally found in clefts or crevices and is dynamic. Substrates are bound to the active site by noncovalent bonds, depending upon the precisely defined arrangement of atoms in the active site. Catalytic residues in the active site participate in bond cleavage and formation during the catalysis.
- *Plasticity and flexibility*: Some enzymes have catalytic residues that are functionally flexible and act upon structurally diverse substrates by different mechanisms (Huber *et al.*, 1994; Kanda *et al.*, 1986; Tsai *et al.*, 1969). The plasticity, in particular the ability of inverting enzymes to use both α- and β-glycosyl substrates, has been described (Matsui *et al.*, 1989; Tanaka *et al.*, 1994). The plasticity of the active sites may also result from different functional groups or the conserved residues at different positions in the primary sequences within the enzyme superfamily to perform the same mechanistic role in the catalysis (Todd *et al.*, 2002).
- *Regulatory enzymes*: Catalytic activities of enzymes are subject to various types of regulations (Hammes, 1982). Enzymes that are subjected to metabolic regulations are commonly referred to as regulatory enzymes.
- *Multienzyme systems and multifunctional enzymes*: Enzymes that mediate sequential reactions in a metabolic pathway, may be organized into multienzyme complexes (Perham, 2000). Enzymes with multiple domains may display more than one enzymatic activity in multifunctional enzymes (Bisswanger and Schmincke-Ott, 1980). The former refers to enzyme systems consist of several polypeptide chains, each having a distinct catalytic activity and associated with one another by noncovalent bonds such as 2-oxoacid dehydrogenase complex (2-oxoacid decarboxylase, lipoate acyltransferase and lipoamide dehydrogenase). The latter refers to enzyme systems with single polypeptide chains having multiple catalytic sites, which are usually located in the separated domains such as ascites hepatoma carbamylphosphate synthetase (carbamylphosphate synthetase, aspartate transcarbamylase, dihydroorotase).
- *Isozymes*: Within a single species, there may exist several different forms of enzyme catalyzing the same reaction, known as isozymes (Markert, 1975). The term isozyme generally refers to those forms of an enzyme that arise from genetically determined differences in amino acid sequences, such as cytoplasmic and mitochondrial malate dehydrogenases. Often isozymes are derived from an association of different subunits in an oligomeric enzyme with different electrophoretic mobilities such as heart and muscle lactate dehyhdogenases.

Molecular structures are available for many enzymes and can be viewed at the Web site maintained by the Biomolecular Structure and Modeling Group at University College in London (http://www.biochem.ucl.ac.uk/bsm/enzymes/index.html)

11.2.2 Catalytic efficiency

As catalysts, enzymes are highly efficient. The catalytic efficiency, in its simplest form, is that an enzyme that operates with maximal efficiency will catalyze the reaction of every

TABLE 11.2 Comparison of catalytic power of various catalysts

Reaction	Catalyst	Temp. (°C)	Rate constant, k ($M^{-1}s^{-1}$)
$C_6H_5CONH_2 + H_2O \rightarrow C_6H_5COOH + NH_4OH$	H^+	52	2.4×10^{-6}
	OH^-	53	8.5×10^{-6}
	α-chymotrypsin	25	14.9
$CO(NH_2)_2 + H_2O \rightarrow CO_2 + 2NH_3$	H^+	62	7.4×10^{-7}
	urease	21	5.0×10^6
$2H_2O_2 \rightarrow 2H_2O + O_2$	Fe^{2+}	22	56
	catalase	22	3.5×10^7

substrate molecule that it encounters. The reaction will be limited by physical steps of diffusion rather than by the chemical transformation (Alberty and Knowles, 1977). Enzyme catalyzed reactions may provide rates of acceleration as much as 10^{12}-fold (Table 11.2).

An enzyme is able to accelerate a chemical reaction by having evolved an active site that is complementary to the transition state. Although an active site must bind substrate, it is in the transition state (relative to all other chemical species in the overall transformation) that the binding interactions with the enzyme are maximized, i.e. $k_{cat}/k_u \approx K_S/K_T$. This illustrates that the enzyme binds the transition-state structure more tightly than it binds the substrate. The relationship between the dissociation constants for the substrate and transition state from the enzyme is of the same order of magnitude as the relationship between the rates of the catalyzed and uncatalyzed reactions (Wolfenden, 1969). Therefore the catalytic efficiency of enzymes is often equated to the free energy of binding of the transition state structures and thus the activation energy differences between the catalyzed and uncatalyzed reactions.

Factors contributing to the catalytic efficiency of enzymes are:

A. Kinetic consequence of complex formation:

1. *Approximation and orientation*: This is the entropic contribution (Page, 1977). Enzymatic reactions take place in the confines of the enzyme-substrate complex. The catalytic groups are part of the same complex as the substrate. This proximity/propinquity and appropriate positioning of substrate molecules with respect to the catalytic groups in the active site via stereopopulation control or orbital steering may contribute to a rate enhancement with a factor of 10 to 10^5.
2. *Substrate anchoring*: The substrate anchoring (Reuben, 1971) may be visualized as having the effect of increasing the probability of forming the activated complex.
3. *Strain and distortion*: The substrate molecule, by forming the enzyme-substrate complex, is forced toward the geometry of the transition state since the enzyme may bind the transition-state-like structure favorably (Secemski and Lienhard, 1971).

B. Catalytic action:

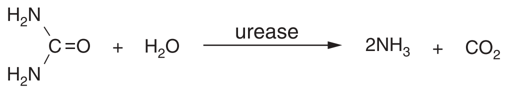
1. *Push-pull catalysis*: Push-pull phenomenon is an effective way in the general acid–base catalysis. Hydrogen bond networks, termed ‘charge relays’ involving substrate molecule and catalytic groups at the active site, have been reported for a number of enzymes (Blow, *et al.*, 1970; Dutler and Brandén, 1981). The network is an effective system for carrying out push-pull catalysis by its hydrogen bonded general acid and base.

2. *Multiple intermediates*: In a single-step conversion, the geometry of the transition state will be some structure between that of the substrate and that of the product, thus having a significantly different geometry with significantly different energy. But in a multiple-step conversion, transition states will not change so much in geometry or energy from those of substrate or product thus having smooth transitions. Furthermore, multiple intermediates avoid freezing the rotational and translational freedoms, thus facilitating the interconversions.
3. *Provision of favorable environment*: The active site of enzymes provides proper microenvironments, which favors the transformation. By bringing in hydrophobic residue(s), the active site can shield against the polar solvent (Pai and Schulz, 1983) to facilitate the reactions involving charge neutralization or dispersion. The perturbations of the active-site environment relative to aqueous solution are used by proteins to increase the energetic contribution to catalysis (Gerlt and Gassman, 1993; Cleland and Kreevoy, 1994). This can be accomplished simply if proteins control their environment to create active sites with effective dielectric constants lower than that in solution. Most protein side chains are hydrophobic, ideal for creating an extensive hydrophobic pocket of low dielectric that favors transition states involving neutralization or dispersion of charges.
4. *Electrostatic stabilization*: In aqueous solutions, a large electrostatic attraction does not arise because any force created by moving opposite charges toward each other is balanced by reorientation of the solvent dipoles. However, enzymes can stabilize varied constellations of charges by fixing their dipoles in appropriate orientations. The active site dipoles associated with polar groups, internal water molecules and ionized residues are already partially oriented toward the transition state charge center, i.e. enzymes use their preorientated environment to stabilize the transition state (Warshell, 1998).

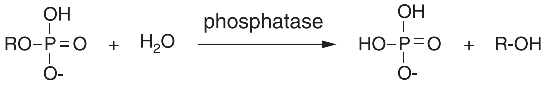
11.2.3 Enzyme specificity

Enzyme specificity refers to the strict limitation of the action of each enzyme to one substance or to a small number of closely related substances. Thus enzyme specificity bestows enzymes with abilities to discriminate among great number of metabolites in the cell, which is the essence of ordered metabolisms of living matter. There are several types of enzyme specificity:

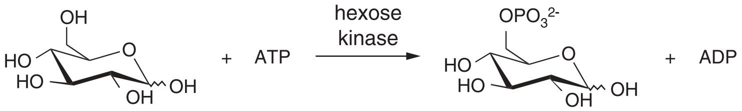
- *Absolute specificity*: Absolute specificity refers to an enzyme that catalyzes only one reaction on only one substrate, such as urease:



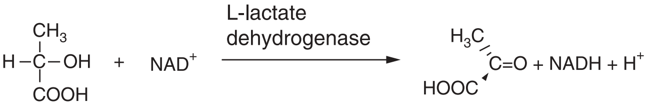
- *Reaction specificity*: Many enzymes promote the same reaction acting on a small number of closely related substances, usually with a particular functional group such as phosphase, which mediates hydrolysis of phosphomonoesters and hexose kinase, which phosphorylates aldohexoses:



$$RO-\overset{\overset{OH}{|}}{\underset{\underset{O^-}{|}}{P}}=O + H_2O \xrightarrow{\text{phosphatase}} HO-\overset{\overset{OH}{|}}{\underset{\underset{O^-}{|}}{P}}=O + R\text{-}OH$$

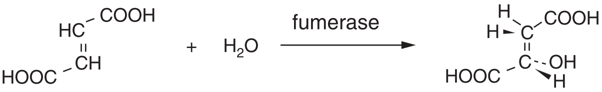


- *Stereospecificity*: Enzymes are also sterically specific when acting on substrates that are stereoisomeric, i.e. isomers in which the atoms are oriented differently in space. The presence of a chiral (asymmetric) center (carbon) in a molecule gives rise to an enatiomeric pair, D and L or R and S (Cahn *et al.*, 1966). Stereospecific enzymes that act on only one enantiomer but not the other are known as chiral stereospecificity. For example, D-lactate dehydrogenase oxidizes only D-lactate to pyruvate and L-lactate dehydrogenase, its L-enantiomer:

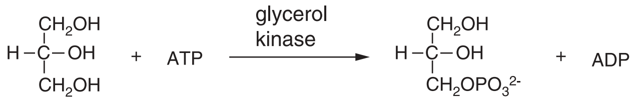


$$\begin{array}{c}CH_3\\ |\\ H-C-OH\\ |\\ COOH\end{array} + NAD^+ \xrightarrow{\text{L-lactate dehydrogenase}} (H_3C)(HOOC)C{=}O + NADH + H^+$$

Stereospecificity may also apply to enzymes acting on the substrates containing double bonds, which give rise to geometric isomers, *cis* and *trans* or E and Z. For example, fumerase specifically catalyzes *trans* hydration of fumerate (*trans*-butenedioic acid):



One of the most subtle stereospecificity of enzymes relates to their ability to distinguish between two identical atoms/groups (proR versus proS) bonded to a carbon atom in prochiral stereospecificity (Hanson and Rose, 1975). For example, yeast fermentative glycerol kinase catalyzes phosphorylation of only one (proS) of the two chemically identical primary hydroxyl groups to yield L-glycerol-3-phosphate:



$$\begin{array}{c}CH_2OH\\ |\\ H-C-OH\\ |\\ CH_2OH\end{array} + ATP \xrightarrow{\text{glycerol kinase}} \begin{array}{c}CH_2OH\\ |\\ H-C-OH\\ |\\ CH_2OPO_3^{2-}\end{array} + ADP$$

According to the three-point attachment model (Ogston, 1948), an enzyme active site has two binding loci, β and γ specific for atoms/groups y and z, and a catalytic locus α specific for the atom/group x of a molecule Cx_2yz. Then this molecule can have only one orientation on the enzyme, which permits the approach of one and only one of the two identical groups x to the catalytic site. In the dissymmetric model (Popják and Cornforth, 1966), a dissymmetric treatment of a substrate by an enzyme can occur wherever the enzyme imposes, whether actively by binding or passively by obstruction, a particular orientation that discriminates the two identical atoms/groups on the substrate at the active site.

11.2.4 Active site of enzyme

The observation that only small portion of a protein molecule is involved in binding substrate and catalytic action lead to the definition of the active site of an enzyme. The active site consists of those amino acid residues (contact amino acids) that are within the bond distance (i.e. 20 nm) from the substrate and those residues (auxiliary amino acids) that play important role in an enzymatic catalysis (Koshland, 1958). Alternatively, the active site is defined as those atoms of side chain residues of proteins directly involved in the catalytic step, i.e. the processes of bond breaking/making it analogous to the catalytic site (Vallee and Riordan, 1969). Those amino acid residues involved directly in the noncovalent attachment of the substrate to an enzyme comprise the (substrate) binding site. The combined catalytic site and binding site is referred to as the active center. Then essential groups are those residues involved in, or in some way required for, a particular property. Therefore their alternation will bring about a loss of the characteristic property of the enzyme (Vallee and Riordan, 1969). The active site of an enzyme has the following attributes:

- The active site of an enzyme is the region that binds the substrates (and cofactors if any) and contributes the catalytic residues that directly participate in the breaking and making of bonds.
- The active site takes up a relatively small part of the total dimension of an enzyme.
- The active site is a three-dimensional entity and dynamic.
- Active sites are generally clefts or crevices.
- Substrates are bound to the active site of enzymes by noncovalent bonds.
- The specificity of binding depends on the precisely defined arrangement of atoms in the active site.

Online information on the catalytic site/active site can be accessed from CSA at http://ebi.ac.uk/thornton-srv/database/CSA/ and PROCAT at http://www.biochem.ucl.ac.uk/bsm/PROCAT/PROCAT.html. SCOPEC (http://www.enzyme.com/database/scopec.php) maps the catalytic function to domain structure of proteins. PDBSum (http://www.ebi.aci.uk/thornton-srv/databases/pdbsum) analyze PDB files for the active/binding site to display its topology (Cleft, Figure 11.1A) and to map the interaction between ligand(s) and amino acid residues at the active/binding site (Ligand/LIGPLOT, Figure 11.1B).

Conceptually, the lock-and-key relationship was invoked to depict the complementarity between lock (enzyme) and key (substrate) at the active site. The induced fit model (Koshland, 1960; Herschlag, 1988) is proposed to explain the action/specificity of an enzyme active site:

- The enzyme exists in a natural conformation, which is not necessary a negative of the substrate.
- The substrate induces a change in protein conformation.
- The changed conformation produces the correct alignment of catalytic groups.
- The interaction of substrate with the active site produces new enzyme-substrate geometry with the lowest energy of activation, which allows the reaction to proceed.
- The release of products from the enzyme allows the protein to revert to its original conformation.

X-ray crystallographic determination of the complexes between enzymes and their substrates (or analogues) or enzymes, and their transition-state analogues provide direct

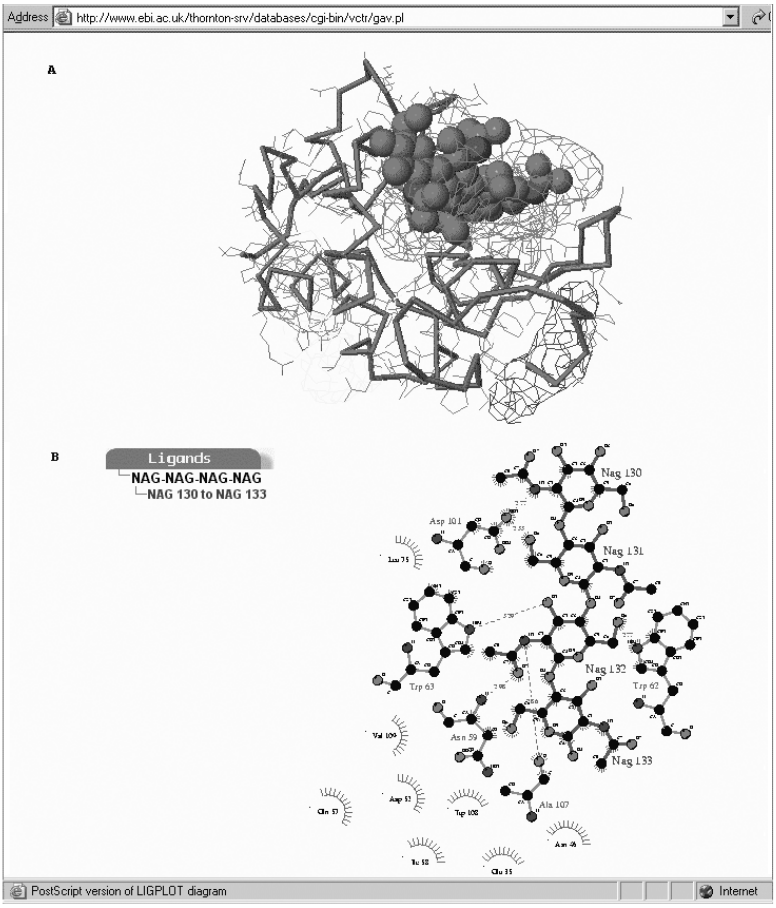

Address
http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/vctr/gav.pl
A
B
Ligands
NAG-NAG-NAG-NAG
NAG 130 to NAG 133
Nag 130
Asp 101
Leu 75
Nag 131
Trp 63
Nag 132
Trp 62
Val 109
Asn 59
Nag 133
Asp 52
Trp 108
Ala 107
Gln 57
Ile 58
Glu 35
PostScript version of LIGPLOT diagram
Internet

Figure 11.1 Active site topology and ligand interaction displayed with PDBSum
The 3D topology of the active site cleft (A) and LIGPLOT (B) for the interaction of N-acetylchitotetraose with the active site residues of lysozyme (1LZC.pdb) are illustrated with PDBSum (http://www.ebi.ac.uk/thornton-srv/databases/pdbsum).

identification of the active site of enzymes (Lolis and Petsko, 1990). Alternative methods, such as chemical modifications (Vallee and Riordan, 1969) site-directed mutagenesis (Smith, 1985) and spectroscopic probes have been employed to investigate the chemical nature of the active sites. These techniques will be discussed in the next section.

11.2.5 Multienzyme complex and multifunctional enzymes

Multienzyme proteins include all proteins with multiple catalytic domains or polypeptide chains. Those which link covalently separate catalytic domains, are generally referred to as multifunctional enzymes or multienzyme polypeptides. Those in which polypeptide chains with different enzymatic activities are organized into protein aggregates, are known as multienzyme complexes. Each of the individual enzymes in an organized multienzyme complex catalyzes its component step in a complex biochemical reaction or one of sequential reactions comprising a segment of a metabolic pathway. For example, *E. coli* pyruvate dehydrogenase complex (PDC), which catalyzes the oxidative decarboxylation of pyruvate to acetyl CoA (Figure 11.2) is composed of three component enzymes (Reed,

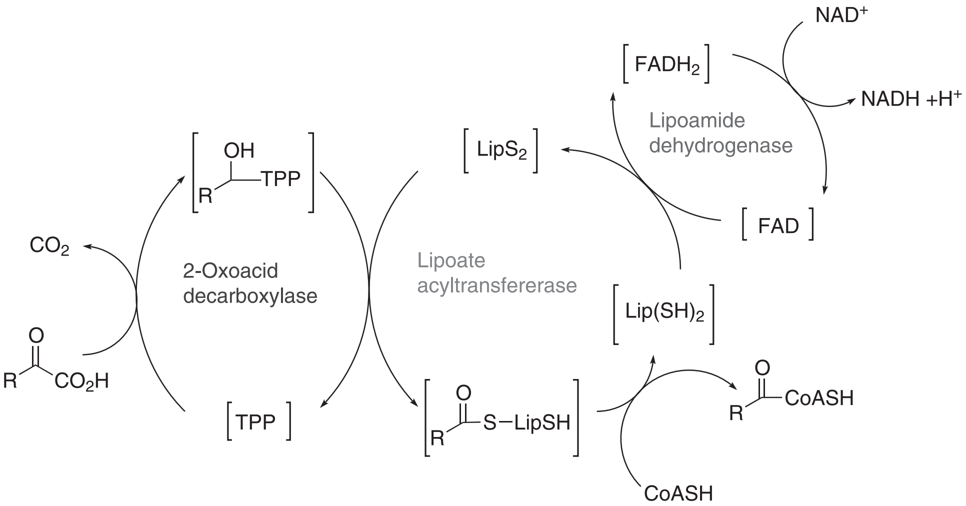


Figure 11.2 Reaction sequences catalyzed by 2-oxoacid dehydrogenase complex
Pyruvate dehydrogenase complex (PDC) and α-ketoglutarate dehydrogenase complex (αKGDC) catalyze the oxidative decarboxylation of pyruvate (R = CH_3) and α-ketoglutarate (R = CH_2CH_2COOH) to Acetyl-CoA and succinyl CoA respectively. Three component enzymes; 2-oxoacid (pyruvate/α-ketoglutarate) decarboxylase, lipoate acetyltransferase/succinyltransferase, dihydrolipoate dehydrogenase as well as five cofactors, namely (1) thiamine pyrophosphate (TPP) and its acylated form, (2) lipoamide ($LipS_2$), reduced form and acylated form, (3) flavin adenine dinucleotide (FAD) and its reduced form, (4) nicotinamide adenine dinucleotide (NAD^+) and its reduced form, and (5) coenzyme A (CoASH) and its acylated product are involved.

1974); pyruvate decarboxylase (E1, a dimer of $M_r = 2 \times 100\,kDa$), lipoate acetyltransferase (E2, $M_r = 80\,kDa$) and lipoamide dehydrogenase (E3, a dimer of $M_r = 2 \times 56\,kDa$).

The complex is constructed around a core of 24 E2 molecules arranged in octahedral symmetry. The lipoate residues that transfer the acetyl group are attached to E2 on ε-NH_2 of Lys. This enables the lipoate to function as a swinging arm of radius 14 Å between E1 (accept acetyl group) and E3 (regenerate oxidized lipoate after delivering the acetyl group to coenzyme A) with a correlation time of 0.2 ns. The mammalian PDC, in addition to the three component enzymes, E1, E2 and E3, also contains two regulatory enzymes, a kinase and a phosphatase. Phosphorylation by the kinase of the complex is accompanied by a proportional decrease in the overall activity, and dephosphorylation by the phosphatase restores the activity. The site of phosphorylation/dephosphorylation is E1, which apparently catalyzes rate-limiting step in the oxidative decarboxylation of pyruvate. Possible advantages of the mutienzyme complexes are:

1. *Catalytic enhancement*: The reduction of diffusion time of an intermediate from one enzyme to the next will enhance the catalytic efficiency of the sequential reactions
2. *Substrate channeling*: The control over an intermediate to follow the designated metabolic route by directing it to a specific enzyme rather than allowing competition from other enzymes in solution.
3. *Sequestration of reactive intermediates*: The protection of chemically unstable intermediates from aqueous solution.
4. *Facilitation of intramolecular transfer reactions*: The proximity of two successive enzymes in the reaction sequence facilitates the direct group transfer.

TABLE 11.3 Examples of moonlighting enzymes

Function A	Additional function(s) B	Switching mechanism
Phosphoglucose mutase	Neuroleukin, autocrine motility factor, differentiation mediator	Interior/exterior of cells: A within the cell cytosol while as B when secreted outside the cell.
Thymidine phosphorylase	Platelet derived endothelial cell growth factor	Interior/exterior of cells: A in cytosol and B in extracellular fluid.
E. coli proline DH/pyrroline-5-carboxylate DH	Transcriptional repressor by DNA binding	Cellular location: A in cytosol and B associated with plasma membrane.
Cytochrome C	Apoptogenic factor	Cellular location: A within mitochondra and B when leaks outside.
Lactate DH, enolase	Lens crystallins	Compartmentation (Tissue): A in cellular cytosol/matochondria whilst B in specialized tissue (eyes).
Glyceraldehyde-3-phosphate DH	Uracil-DNA glycosylase	Oligomerization: A in tetramer but B in monomer.
Aconitase	Iron-responsible binding protein	Ligand concentration: A in high iron conc. but B in decreased iron conc.
Carbinolamine dehydratase	Dimerization cofactor of HTF and HNF-1α	Complex formation: A in uncomplexed but B in complexed forms.

Note: Abbreviations used are: DH, dehydrogenase; HTF, homeodomain transcription factor; HNF, hepatic nuclear factor.

Enzymes that combine several autonomous functions on one polypeptide chain are termed multifunctional enzymes. Autonomous function is normally taken as that each function is assigned to a distinct region (i.e. a domain) on the polypeptide chain, such as aspartokinae (/homoserine dehydrogenase), Chorismate mutase (/prephenate dehydratase), DNA polymerase I (/exonucleases) of *E. coli*. There is an interesting category of multifunctional enzymes known as multitasking/moonlighting enzymes (Jeffery, 2004), which display a major function that can be switched to additional function(s) (Table 11.3).

11.3 ENZYME KINETICS

11.3.1 Fundamental of enzyme kinetics

Enzymes are biocatalysts, which facilitate rates of biochemical reactions. Enzyme kinetics (Cornish-Bowden, 1995; Marangoni, 2003; Plowman, 1972; Segel, 1975; Schulz, 1994) is a detailed step-wise study of enzyme catalysis as affected by various factors. The most important factors that affect the rates of enzymatic reactions are enzyme concentration, ligand (substrates, products, inhibitors and activators) concentrations, solvent (solution, ionic strength and pH) and temperature. When all these factors are properly analyzed, it is possible to learn a great deal about the nature of enzymes. Rates of an enzymatic reaction (v) are normally measured in which $v = kE_0$, where E_0 represents the total (initial) enzyme concentration and detailed kinetic studies are performed by varying concentrations of ligand(s) to obtain kinetic parameters, e.g. maximum velocity (V), Michaelis–Menten constant (K_m or K_a) and/or inhibition constant (K_i). These kinetic parameters are essential in an understanding of the kinetic mechanism of the reaction, i.e. the order in which the substrates add and products leave or the manner in which the inhibitor acts.

Basic approaches to solve enzyme kinetics follow:

1. S*ystem*: A simplified reaction involving one substrate (uni-substrate enzyme catalysis) for the forward direction as shown by:

$$E + A \underset{k_2}{\overset{k_1}{\rightleftarrows}} EA \xrightarrow{k_3} E + P$$

   where E, A, EA and P are free enzyme, substrate, enzyme-substrate complex and product respectively. k_1, k_2 and k_3 are respective rate constants.

2. *Condition*: Initial velocities (v) are generally measured at a very low enzyme concentration (i.e. 10^{-8}–10^{-10} M) compared with substrate concentrations (A), which are usually greater than 10^{-6} M. Thus

$$A_0 \gg E_0$$

3. *Conservation equations*: Two conservation equations can be written as

$$E = E_0 + EA$$
$$A = A_0 + EA \approx A_0$$

4. *Initial velocity expression*: Identification of the rate-determining step and description of the initial velocity expression:

$$v = k_3EA$$

5. *Kinetic treatments*: Enzyme kinetic data are treated according to two assumptions:

   A. *Quasi-equilibrium assumption* also known as the rapid-equilibrium assumption in which an equilibrium condition exists between the enzyme E, its substrate (A) and the enzyme-substrate complex (EA), i.e.:

$$K_m = k_2/k_1 = (E)(A)/(EA) \approx (E_0 + EA)(A_0)/(EA)$$

   Substitute $(EA) = (E_0)(A_0)/(A_0 + K_m)$ into the initial velocity expression and let $V = k_3E_0$, i.e. the maximum velocity is reached when all the enzyme molecules are complexed with substrate molecules:

$$v = k_3(E_0)(A_0)/(A_0 + K_m) = V \cdot A/(A + K_m)$$

   This is the Michaelis–Menten equation, where V and K_m are the maximum velocity and Michaelis–Menten constant (sometime simply termed Michaelis constant).

   B. *Steady-state assumption* in which the steady state concentration of the enzyme-substrate complex is reached/maintained by the rates of its formation and decomposition (the concentration of the enzyme-substrate complex does not change), i.e.:

$$d(EA)/dt = k_1(E)(A) - k_2(EA) - k_3(EA) = k_1(E_0)(A_0) - (k_1A_t + k_2 + k_3)(EA) = 0$$

   Substitute $(EA) = k_1(E_0)(A_0)/(k_1A_t + k_2 + k_3)$ into the initial velocity expression and let $K_a = (k_2 + k_3)/k_1$ and $V = k_3E_t$:

$$v = k_3(E_0)(A_0)/[A_0 + \{(k_2 + k_3)/k_1\}] = V \cdot A/(A + K_a)$$

6. *Commonality*: Both assumptions yield rate equations with the identical form, except with different constants in which $K_m = k_2/k_1$ according to the quasi-equilibrium and $K_a = (k_2 + k_3)/k_1$ based on the steady-state. If indeed k_3 is rate-limiting and therefore

$k_3 \ll k_2$, then $K_m = K_a$. In the subsequent discussion of enzyme kinetics, K_a will be used in accordance with Cleland nomenclature (Cleland, 1963):

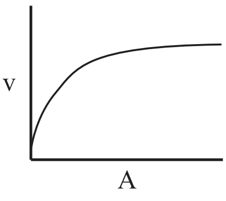
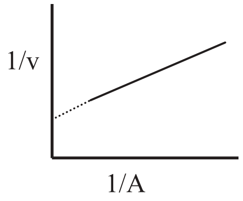
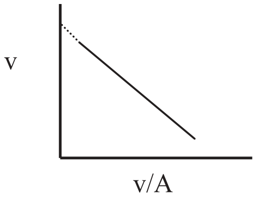
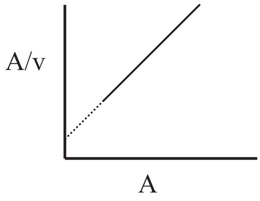
7. *Implications of kinetic parameters*: Kinetic parameters have the following meanings:
   a) *Turnover number*, k_{cat} is defined as $k_{cat} = V/E_0$ that measures the efficiency of an enzyme because it measures how rapidly (efficiency) an enzyme can operate once the active site is completely occupied (at high substrate concentrations).
   b) *Michaelis–Menten constant* equates to the dissociation constant of the enzyme-substrate complex, thus it measures the affinity of an enzyme, i.e. the smaller the constant, the higher the affinity.
   c) *Specificity constant* is defined as k_{cat}/K_a (or k_{cat}/K_m) that compares the relative ability of different compounds to serve as substrates for the same enzyme. Under physiological conditions operated at low substrate concentrations, $V/K_a = v/A$ (If $A \ll K_a$, $v \approx VA/K_a$ and $Ka = VA/v$, therefore $V/Ka = v/A$), thus $k_{cat}/K_a = (v/A)/E_0$ measures velocity of catalysis for the substrate by the enzyme.

The substrate saturation curve (substrate concentrations versus initial velocities) is hyperbolic and its asymptote gives an estimate of the maximum velocity while the substrate concentration at the half-maximum velocity provides the Michaelis–Menten constant. Historically these kinetic parameters can be obtained readily from the linearly transformed Michaelis–Menten equations (Table 11.4). However, these parameters are now routinely evaluated by the statistical and computer analyses (Wilkinson, 1961; Cleland, 1967; Cornish-Bowden, 1995).

The foregoing simplified, uni-substrate enzyme catalysis that is presented to convey the concept of enzyme kinetics, needs to be modified in order to describe the majority of enzyme reactions that are reversible, such as:

$$E + A \underset{k_2}{\overset{k_1}{\rightleftarrows}} EA \underset{k_4}{\overset{k_3}{\rightleftarrows}} EP \underset{k_6}{\overset{k_5}{\rightleftarrows}} E + P$$

TABLE 11.4 Linear transformations of Michaelis-Menten equation

	Substrate saturation	Lineweaver-Burk	Eadie	Hofstee
Eqn.	$V = VA/(A + K_a)$	$1/v = 1/V + (K_a/V)(1/A)$	$v = V - K_a(v/A)$	$A/v = K_a/V + (1/V)A$
Plot		1/v versus 1/A	V versus v/A	A/V versus A
				
V =	Asymptote	1/Intercept	Intercept	1/Slope
K_a =	[A] at $^1/_2$V	Slope/Intercept	– Slope	Intercept/Slope

Notes: It is recommended that statistical and computer analysis of kinetic data should be carried out to evaluate kinetic parameters (Cleland, 1967; Cornish-Bowden, 1995), however these linear plots may still retain their diagnostic values. The linear plots indicate the compliance to the Michaelis–Menten kinetics whereas nonlinear plots imply multiple substrate addition, substrate inhibition or homotropic allosterism.

The reversible reaction involves an interconversion of the two enzyme complexes, namely the enzyme-substrate (EA) and the enzyme-product (EP) complexes. Therefore the steady-state assumption considers the steady-state concentrations of these two complexes, dEA/dt = 0 and dEP/dt = 0. The kinetic treatment gives a rate equation in the Cleland form:

$$v = \frac{V_1V_2(A - P/K_{eq})}{V_2K_a + V_2A + (V_1P)/K_{eq}}$$

where maximum velocities, $V_1 = k_3k_5Et/(k_3 + k_4 + k_5)$ for the forward direction and $V_2 = k_2k_4E_t/(k_2 + k_3 + k_4)$ for the reverse direction. The Haldanes relationship for the equilibrium constant, $K_{eq} = (k_1k_3k_5)/(k_2k_4k_6) = (V_1K_p)/(V_2K_a)$ where Michaelis–Menten constants, $K_a = (k_2k_4 + k_2k_5 + k_3k_5)/\{k_1(k_3 + k_4 + k_5)\}$ for the forward direction and $K_p = (k_2k_4 + k_2k_5 + k_3k_5)/\{k_6(k_2 + k_3 + k_4)\}$ for the reverse direction.

The quasi-equilibrium treatment yields an identical rate equation if represented in the Cleland form. However, the quasi-equilibrium condition assumes k_3 and k_4 being rate-limiting, which yields different expressions for the maximum velocities, $V_1 = k_3Et$, $V_2 = k_4Et$ and Michaelis-Menten constants, $K_a = k_2/k_1$, $K_p = k_6/k_5$ respectively.

Although the steady state treatment is the preferred approach for analyzing enzyme kinetic data, the applications of both kinetic treatments in general enzyme reactions will be considered.

11.3.2 Steady-state kinetic treatment of enzyme catalysis

The steady-state treatment of enzyme kinetics assumes that concentrations of the enzyme-containing intermediates remain constant during the period over which an initial velocity of the reaction is measured. Thus the rates of changes in the concentrations of all the enzyme-containing species equal zero. Under the same experimental conditions, i.e. $A_0 \gg E_0$, the general rule for writing the rate equation according to the King and Altman method (King and Altman, 1956) can be illustrated for the sequential bisubstrate reaction. Figure 11.3 shows the ordered bi bi reaction with substrates A and B forming products P and Q. Each enzyme-containing species is associated with two rate constants, k_{odd} in the forward direction and k_{even} in the reverse direction.

The steady-state rate equation is obtained according to following rules (King and Altman method) if no loop is involved:

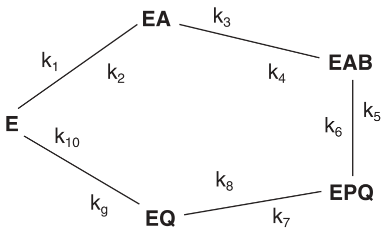
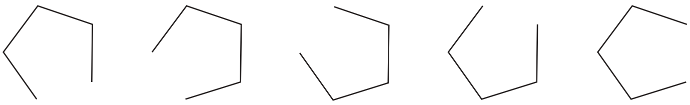


Figure 11.3 Diagram for ordered bi bi kinetic mechanism
The free enzyme, E binds to A (first substrate) to form binary complex, EA which then interacts with B (second substrate) to form ternary complex, EAB. The two ternary complexes EAB and EPQ interconvert. The release of P (first product) forms EQ which then dissociates to E and Q (second product) in an ordered sequence. k_{1-10} are rate constants.

1. Write down all possible basic pattern with n − 1 lines (n = number of enzyme forms i.e. enzyme-containing species), in which all lines are connected without closed loops e.g.:



2. The total line patterns equals m!/{(n − 1)!(m − n + 1)!} where m is the number of lines in the patterns.
3. Write a distribution equation for each enzyme form, e.g., $E/E_0 = N_e/D$ where N_e and D are the numerator (for E) and denominator terms respectively.
4. Numerator terms (e.g. N_e) are written:
    (a) Follow along the lines in the basic pattern in the direction from other enzyme forms (i.e. EA, EQ, EAB and EPQ) leading toward the free enzyme form (i.e. E) for which the numerator is sought.
    (b) Multiply all the rate constants and concentration factors for this direction.
    (c) Repeat the process for all the basic patterns.
    (d) The numerator terms are the sum of all the products of rate constants and concentration factors.
5. Write the numerator terms for all the other enzyme forms by repeating the process e.g. N_{ea}, N_{eq}, N_{eab} and N_{epq} for EA/E_0, EQ/E_0, EAB/E_0, and EPQ/E_0 respectively.
6. Denominator terms are the sum of all numerator terms i.e. $D = N_e + N_{ea} + N_{eq} + N_{eab} + N_{epq}$.
7. Substitute appropriate distribution equations (*e.g.* EPQ/E_0 and EQ/E_0) into the initial velocity expression:

$$
\begin{aligned}
v &= dP/dt = k_7[EPQ] - k_8[EQ][P] = k_7\{EPQ/E_0\}[E]_0 - k_8\{EQ/E_0\}[E]_0[P] \\
&= (k_1k_3k_5k_7k_9[A][B] - k_2k_4k_6k_8k_{10}[P][Q])[E]_0)/\{k_2(k_4k_6 + k_4k_7 + k_5k_7)k_9 \\
&\quad + k_1(k_4k_6 + k_4k_7 + k_5k_7)k_9[A] + k_3k_5k_7k_9[B] + k_2k_4k_6k_8[P] + k_1k_3(k_5k_7 + k_5k_9 \\
&\quad + k_6k_9 + k_7k_9)[A][B] + (k_2k_4 + k_2k_5 + k_2k_6 + k_4k_6)k_8k_{10}[P][Q] + k_1k_4k_6k_8[A][P] \\
&\quad + k_3k_5k_7k_{10}[B][Q] + k_1k_3k_8(k_5 + k_6)[A][B][P] + k_3k_8k_{10}(k_5 + k_6)[B][P][Q]
\end{aligned}
$$

This steady-state equation expressed with rate constants can be converted into the rate equation expressed with kinetic parameters according to Cleland (Cleland, 1963):

$$
v = \frac{V_1V_2(AB - PQ/K_{eq})}{\begin{array}{c}(K_{ia}K_b + K_bA + K_aB + AB)V_2 + (K_pQ + K_qP + PQ)V_1/K_{eq} \\ +(K_aBQ/K_{iq} + ABP/K_{ip})V_2 + (K_qAP/K_{ia} + BPQ/K_{ib})V_1/K_{eq}\end{array}}
$$

where the equilibrium constant, $K_{eq} = (V_1K_pK_{iq})/(V_2K_bK_{ia})$. V_1 and V_2 are maximum velocities for the forward and reverse reactions. K_a, K_b, K_p and K_q are Michaelis constants, while K_{ia}, K_{ib}, K_{ip} and K_{iq} are inhibition constants associated with substrates (A and B) and products (P and Q) respectively. The rate equation for the forward reaction can be simplified (Table 11.4) to

$$
v = \frac{V_1AB}{K_{ia}K_b^{`} + K_bA + K_aB + AB}
$$

11.3.3 Quasi-equilibrium treatment of random reactions

The steady-state kinetic treatment of random reactions is complex and gives rise to rate equations of higher order in substrate and product terms. For kinetic treatment of random reactions that display the Michaelis–Menten (i.e. hyperbolic velocity-substrate relationship) or linear (linearly transformed kinetic plots) kinetic behavior, the quasi-equilibrium assumption is commonly made to analyze enzyme kinetic data.

The general rule for writing the rate equation according to the quasi-equilibrium treatment of enzyme kinetics can be exemplified for the random bisubstrate reaction with substrates A and B forming products P and Q (Figure 11.4), where $K_aK_{ab} = K_bK_{ba}$ and $K_pK_{pq} = K_qK_{qp}$.

1. Write the initial velocity expression: $v = k[EAB] - k'[EPQ]$ where the interconversion between the ternary complexes is associated with the rate constants, k and k′ in the forward and reverse directions respectively.
2. Divide the velocity expression by the conservation equation for enzyme, $[E]_0 = [E] + [EA] + [EB] + [EAB] + [EPQ] + [EP] + [EQ]$, i.e.

   $v/[E]_0 = (k[EAB] - k'[EPQ])/([E] + [EA] + [EB] + [EAB] + [EPQ] + [EP] + [EQ])$.
3. Set $[E]_0$ term equal unity, i.e. $[E]_0 = 1$.
4. The term for any enzyme-containing complex is composed of a numerator, which is the product of the concentrations of all ligands in the complex and a denominator, which is the product of all dissociation constants between the complex and free enzyme, i.e. $[EA] = [A]/K_a$, $[EB] = [B]/K_b$, $[EP] = [P]/K_p$, $[EQ] = [Q]/K_q$, $[EAB] = ([A][B])/(K_aK_{ab})$, and $[EPQ] = ([P][Q])/(K_qK_{qp})$.
5. Substitution yields the rate expression:

$$v = \frac{\{k([A][B])/(K_aK_{ab}) - k'([P][Q])/(K_qK_{qp})\}[E]_0}{1`+[A]/K_a + [B]/K_b + ([A][B])/(K_aK_{ab}) + [P]/K_p + [Q]/K_q + ([P][Q])/(K_qK_{qp})}$$

The rate expression for the forward direction is simplify to

$$v = \frac{kAB/(K_aK_{ab})E_0}{1`+A/K_a + B/K_b + AB/(K_aK_{ab})} = \frac{VAB}{K_aK_{ab}`+K_{ab}A + K_{ba}B + AB}$$

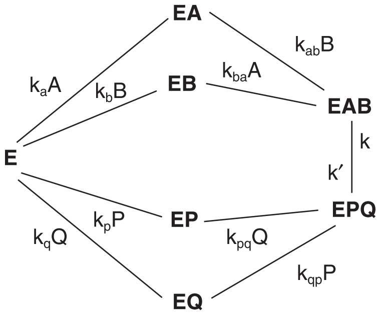


Figure 11.4 Diagram for random bi bi kinetic mechanism
The random addition of substrates, A and B forms binary (EA and EB) and ternary (EAB) complexes. The two ternary complexes, EAB and EPQ interconvert with the rate constant of k and k′. The release of products, P and Q also proceeds in a random manner *via* binary complexes (EP and EQ). K's are dissociation constants where $K_aK_{ab} = K_bK_{ba}$ and $K_pK_{pq} = K_qK_{qp}$.

11.3.4 Cleland's approach

The Cleland nomenclature (Cleland, 1963) for enzyme reactions follows:

1. The number of kinetically important substrates or products is designated by the syllables, *Uni, Bi, Ter, Quad, Pent etc*. as they appear in the mechanism.
2. A *sequential* mechanism will be one in which all the substrates must be present on the enzyme before any product can leave. Sequential mechanisms will be designated as ordered or random, depending on whether the substrate adds and the product releases in an obligatory sequence or in a nonobligatory sequence.
3. A *ping pong* mechanism will be designated if one or more products are released during the substrate addition sequence, thereby breaking the substrate addition sequence into two or more segments. Each segment is given an appropriate syllable corresponding to the number of substrate additions and product releases.
4. The letters, A, B, C, D designate substrates in the order of their addition to the enzyme. Products are P, Q, R, S in the order of their release. Stable enzyme forms are designated by E, F, G, H with E being free enzyme.
5. Isomerization of a stable enzyme formed as a part of the reaction sequence is designated by the prefix *Iso*, such as *Iso ordered*, *Iso ping pong*.
6. For expressing enzymatic reactions, the sequence is written from left to right with a horizontal line or group of lines representing the enzyme in its various forms. Substrate additions and product releases are indicated by the downward (↓) and upward (↑) vertical arrows respectively.
7. Each arrow is associated with the corresponding reversible step, i.e. one rate constant each with the forward and reverse directions. Generally, odd-numbered rate constants are used for the forward reactions whereas even-numbered ones are used for the reverse direction.

Examples of the bisubstrate reactions according to Cleland nomenclature are listed in Table 11.5.

11.3.5 Nonlinear kinetics

The steady-state kinetic treatment of multisubstrate random enzyme reactions gives rise to the forward rate equation of higher order in substrate terms that reflect the number of substrate addition in the formation of intermediary complexes. The transformations are nonlinear. For example, the steady-state treatment of the random bi bi reaction gives, in a coefficient form:

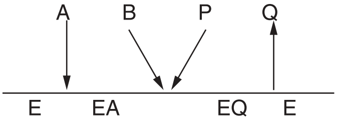
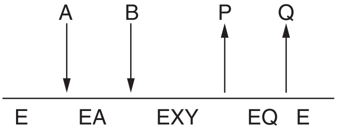
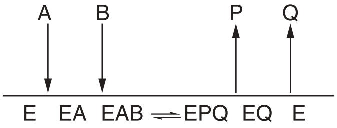
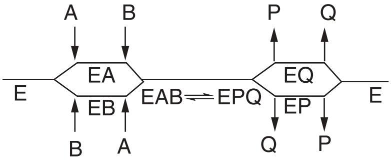
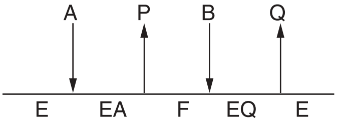
$$v = \frac{\{(n_1B + n_2B^2)A + (n_3 + n_4B^2)A^2\}E_0}{(d_0 + d_1B + 2B^2) + (d_3 + d_4B + d_5B^2)A + (d_6 + d_7B + d_8B^2)A^2}$$

where the n's and d's are numerator and denominator coefficients consisting of composites of rate constants. The equation can be further simplified by keeping one of the substrates constant:

$$v/E_0 = (\alpha_1A + \alpha_2A^2)/(\beta_0 + \beta_1A + \beta_2A^2)$$

where α's and β's are numerator and denominator terms, incorporating the constant B into respective coefficients. The rate equation for the random multisubstrate enzyme reaction can be expressed with respect to each substrate as an n : m function, where n and m are

TABLE 11.5 Cleland nomenclature for bisubstrate reactions exemplified. Three common kinetic mechanisms for bisubstrate enzymatic reactions are exemplified. The forward rate equations for the order bi bi and ping pong bi bi are derived according to the steady-state assumption, whereas that of the random bi bi is based on the quasi-equilibrium assumption. These rate equations are first order in both A and B, and their double reciprocal plots (1/v *versus* 1/A or 1/B) are linear. They are convergent for the order bi bi and random bi bi but parallel for the ping pong bi bi due to the absence of the constant term ($K_{ia}K_b$) in the denominator. These three kinetic mechanisms can be further differentiated by their product inhibition patterns (Cleland, 1963b)

Mechanism	Cleland representation	Forward rate equation
Order bi bi: No ternary complex		$v = \frac{V_1AB}{K_{ia}K_b + K_bA + K_aB + AB}$
One ternary complex		
Two ternary complex		
Random bi bi	 A B P Q E EA EB EAB ⇌ EPQ EQ EP E B A Q P	$v = \frac{V_1AB}{K_{ia}K_b + K_bA + K_aB + AB}$
Ping pong bi bi or Uni uni uni uni ping pong		$v = \frac{V_1AB}{K_bA + K_aB + AB}$

Notes: 1. V_1, K_a, K_b and K_{ia} are maximum velocity, Michaelis-Menten constants for A and B, and inhibition constant for A respectively. 2. The Order bi bi reactions may display zero (known as Theorell-Chance mechanism), one and two ternary complexes. All of them give the rate equation in the same Cleland form.

the highest order of the substrate for the numerator and denominator terms respectively (Bardsley and Childs, 1975). Thus the forward rate equation for the random bi bi derived according to the quasi-equilibrium assumption is 1 : 1 function in both A and B (i.e. first order in both A and B). However, the rate equation for the random bi bi based on the steady-state assumption yields 2 : 2 function (i.e. second order in both A and B). The double reciprocal transformation of the 2 : 2 function gives a plot, which is linear in the asymptotic region but nonlinear in the concave region (Figure 11.5).

The nonlinear 2 : 2 function kinetics should be differentiated from other nonlinear kinetics such as allosteric/cooperative kinetics (Bardsley and Waight, 1978) and the formation of the abortive substrate complex (Dalziel and Dickinson, 1966). The cooperative kinetics (of the double reciprocal plots) can either concave up (positive cooperativity) or

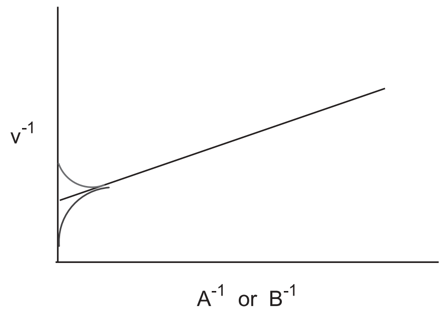


Figure 11.5 Double reciprocal plot of steady state random bi bi enzyme reactions
The double reciprocal transformation of a 2:2 function in A gives: $E_0/v = (\beta_2 + \beta_1 A^{-1} + \beta_0 A^{-2})/(\alpha_2 + \alpha_1 A^{-1}) = \beta_1/\alpha_1 - \alpha_2\beta_0/\alpha_1^2 + \beta_0/\alpha_1(A^{-1}) + (\alpha_1^2\beta_2 + \alpha_2^2\beta_0 - \alpha_1\alpha_2\beta_1)/[\alpha_1^2\{\alpha_2 + \alpha_1(A^{-1})\}]$ Thus, the plot is (1) linear if $\alpha_1^2\beta_2 + \alpha_2^2\beta_0 = \alpha_1\alpha_2\beta_1$ (black), (2) concave up if $\alpha_1^2\beta_2 + \alpha_2^2\beta_0 > \alpha_1\alpha_2\beta_1$ (red), and (3) concave down if $\alpha_1^2\beta_2 + \alpha_2^2\beta_0 < \alpha_1\alpha_2\beta_1$ (blue).

concave down (negative cooperativity), whereas the abortive substrate complex results in a concave-up curve (of the double reciprocal plots) approaching the infinite toward the 1/v-axis (Tsai, 1978).

11.3.6 Environmental effects

The rate of an enzymatic reaction is affected by a number of environmental factors; such as solvent, ionic strength, temperature, pH and presence of inhibitor/activator. Some of these effects are described below:

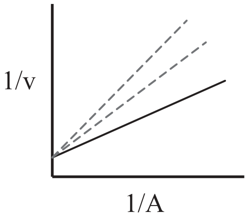
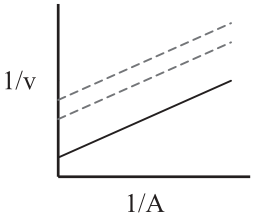
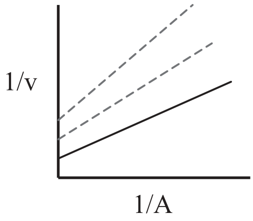
11.3.6.1 **Presence of inhibitors*: inhibition kinetics.*** The kinetic study of an enzymatic reaction in the presence of inhibitors is one of the most important diagnostic procedures for enzymologists. The inhibition (reduction in the rate) of an enzyme reaction is one of the major regulatory devices of living cells and offers great potential for the development of pharmaceuticals. An irreversible inhibitor forms the stable enzyme complex or modifies the enzyme to abolish its activity, whereas a reversible inhibitor (I) forms dynamic complex(es) in equilibrium with the enzyme (E) or the enzyme substrate complex (EA), by reducing the rate of the enzymatic reaction:

$$\begin{tikzcd}
E + A \arrow[r, shift left] & EA \arrow[l, shift left] \arrow[r] & E + P \\
+ & + & \\
I \arrow[d, shift right] & I \arrow[d, shift right] & \\
EI \arrow[u, shift right] & EAI \arrow[u, shift right] &
\end{tikzcd}$$

It is assumed that the inhibitor complex EI cannot combine with the substrate and the EAI complex cannot be decomposed to form the product, therefore establishes an equilibria for the formation of the inhibitor complexes.

The general rule for writing the rate equation according to the equilibrium treatment of enzyme inhibition kinetics for the forward direction is as follows:

TABLE 11.6 Reversible inhibition patterns of enzyme reactions

Type of inhibition	Competitive	Uncompetitive	Noncompetitive
Inhibitor complex	$E + I \xleftrightarrow{K_{is}} EI$	$EA + I \xleftrightarrow{K_{ii}} EAI$	$EA + I \xleftrightarrow{K_{is}} EI$ $EA + I \xleftrightarrow{K_{ii}} EAI$
Rate equation	$v = \dfrac{V_1 A}{K_a(1 + I/K_{is}) + A}$	$V = \dfrac{V_1 A}{K_a + A(1 + I/K_{ii})}$	$v = \dfrac{V_1 A}{K_a(1 + I/K_{is}) + A(1 + I/K_{ii})}$
Effect	Slope effect	Intercept effect	Slope and intercept effects
Double reciprocal plot in the presence of I (red dashed lines)			

Notes: K_{ii} and K_{is} are inhibition (dissociation) constants for the formation of the inhibitor complexes, EI and EAI respectively. Noncompetitive inhibition is also known as mixed competitive inhibition.

1. Write the initial velocity expression, v = k[EA], where k is the rate constant in the forward direction.
2. Divide the velocity expression by the conservation equation for enzyme, $[E]_0 = [E] + [EA] + [EI] + [EAI]$, i.e. $v/[E]_0 = k[EA]/([E] + [EA] + [EI] + [EAI])$.
3. Set [E] term equal unity, i.e. [E] = 1.
4. The term for any enzyme-containing complex is composed of a numerator, which is the product of concentrations of all ligands in the complex and a denominator, which is the product of all dissociation constants between the complex and the free enzyme, i.e. $[EA] = [A]/K_a$, $[EI] = [I]/K_{is}$, and $[EAI] = ([A][I])/(K_aK_{ii})$. The subscripts *is* and *ii* denote slope effect and intercept effect respectively of the double reciprocal velocity plots.
5. The substitution yields the rate expression:

$$v = \frac{(k[A]/K_a)[E]_0}{1 + [A]/K_a + [I]/K_{is} + ([A][I])/K_aK_{ii}} = \frac{V[A]}{K_a(1 + [I]/K_{is}) + [A](1 + [I]/K_{ii})}$$

This is the mixed competitive (noncompetitive) inhibition in which I combines with both E and EA. The inhibition is competitive if I combines only with E, whereas it is uncompetitive if I combines only with EA (Table 11.6).

The three types of reversible inhibitions can be differentiated by double reciprocal rate plots, that:

1. the competitive inhibitors affect only slopes (K_{is} for the slope effect);
2. the uncompetitive inhibitors affect only the intercepts (K_{ii} for the intercept effect); and
3. the noncompetitive inhibitors affect both slopes and intercepts.

The double reciprocal plots are linear (linear inhibition) only when the inhibitor forms the completely inactive complex with one enzyme species (the inhibitor complex formation can be treated by an equilibrium assumption and the complex does not yield the product). Otherwise a nonlinear inhibition may result if the inhibitor complex is partially active or if the multiple inhibitor complexes are formed with an enzyme species or with multiple enzyme species.

11.3.6.2 **Temperature effect*: determination of activation energy.***

From the transition state theory of chemical reactions, an expression for the variation of the rate constant, k with temperature, known as the Arrhenius equation, can be written as

$$k = Ae^{-Ea/RT}$$

or

$$\ln k = \ln A - Ea/RT$$

where A, R and T are pre-exponential factor (collision frequency), gas constant, and absolute temperature respectively. Ea is the activation energy that is related to the enthalpy of formation of the transition state complex, $\Delta H^{\ddagger}$ of the reaction ($Ea = \Delta H^{\ddagger} + RT$). The lowering of the activation energy of an enzymatic reaction is achieved by the introduction into the reaction pathway of a number of reaction intermediate(s).

11.3.6.3 **pH effect*: estimation of pKa value(s).***

Some of the possible effects that are caused by a change in pH are:

1. change in the ionization of groups involved in catalysis;
2. change in the ionization of groups involved in binding the substrate;
3. change in the ionization of substrate(s);
4. change in the ionization of other groups in the enzyme; and
5. denaturation of the enzyme.

The pH effect on kinetic parameters (pH-rate/binding profile) may provide useful information on the ionizing groups of the enzyme if the kinetic studies are carried out with an nonionizable substrate in the pH region (pH 5–9) where enzyme denaturation is minimum. If the Michaelis constant (K) and/or the maximum velocity (V) vary with pH, the number and pK values of the ionizing group(s) can be inferred from the shape of pH-rate profile (pH vs. pK and pH vs. logV plots), namely full bell shape for two ionizing groups and half bell shape for one ionizing group, as summarized in Table 11.7.

TABLE 11.7 pH effects on enzyme kinetics

		Diagnostic	
pH-Rate profile	Rate Expression	Low pH	High pH
Full bell	$K = K_m/(H/K_{1e} + 1 + K_{2e}/H)$	$\log K = \log K_m - pK_{1e} + pH$	$\log K = \log K_m + pK_{2e} - pH$
Left half bell	$K = K_m/(H/K_{1e} + 1)$	$\log K = \log K_m - pK_{1e} + pH$	
Right half bell	$K = K_m/(1 + K_{2e}/H)$		$\log K = \log K_m + pK_{2e} - pH$
Full bell	$V = V_m/(H/K_{1es} + 1 + K_{2es}/H)$	$\log V = \log V_m - pK_{1es} + pH$	$\log V = \log V_m + pK_{2es} - pH$
Left half bell	$V = V_m/(H/K_{1es} + 1)$	$\log V = \log V_m - pK_{1es} + pH$	
Right half bell	$V = V_m/(1 + K_{2es}/H)$		$\log V = \log V_m + pK_{2es} - pH$

Note: K_{1e} and K_{2e} or K_{1es} and K_{2es} are ionizing group(s) in the free enzyme or the enzyme-substrate complex respectively.

The initial rate enzyme kinetics uses very low enzyme concentrations (e.g. 0.1 μM to 0.1 pM) to investigate the steady-state region of enzyme catalyzed reactions. To investigate an enzymatic reaction before the steady-state (i.e. transient state), special techniques known as transient kinetics (Eigen and Hammes, 1963) are employed. The student should consult chapters of kinetic texts (Hammes, 1982; Robert, 1977) on the topics. KinTekSim (http://www.kintek-corp.com/kinteksim.htm) is the Windows version of KINSIM/FITSIM (Frieden, 1993), which analyzes and simulates enzyme catalyzed reactions.

11.4 ENZYME MECHANISMS

11.4.1 Essay on enzyme reaction mechanism

An enzyme reaction mechanism can be studied and understood in three different levels (Lolis and Petsko, 1990):

1. *Kinetic mechanism*: The order of binding of substrates and dissociation of products, and kinetic parameter such as k_{cat} and K_a, are determined.
2. *Chemical mechanism*: The various chemical species, such as intermediates and transition states that are formed during the course of a reaction, are defined. In conjunction with kinetic mechanism, free energy profiles, which describes the energetics of chemical pathway, can be obtained.
3. *Structural mechanism*: The study seeks to understand how the enzyme lowers the activation energy of the reaction and is able to nudge the substrate along the pathway that leads to a product.

A variety of techniques have been applied to investigate enzyme reaction mechanisms. Kinetic and X-ray crystallographic studies have made major contributions to the elucidation of enzyme mechanisms. Valuable information has been gained from chemical, spectroscopic and biochemical studies of the transition-state structures and intermediates of enzyme catalysis. Computational studies provide necessary refinement toward our understanding of enzyme mechanisms. The ability of an enzyme to accelerate the rate of a chemical reaction derives from the complementarity of the enzyme's active site structure to the activated complex. The transition state by definition has a very short lifetime ($\sim 10^{-12}$ s). Stabilization of the transition state alone is necessary but not sufficient to give catalysis, which requires differential binding of substrate and transition state. Thus a detailed enzyme reaction mechanism can be proposed only when kinetic, chemical and structural components have been studied. The online enzyme catalytic mechanism database is accessible at EzCatDB (http://mbs.cbrc.jp/EzCatDB/).

The formulated mechanism must not only be in compliance with chemical principles of catalysis and structural characteristics of proteins, but also be able to explain catalytic efficiency and specificity of enzymatic reactions (Fersht, 1985; Walsh, 1979). Mechanistically the rate enhancement by enzymes is achieved by:

A. *Noncovalent catalysis*. The catalytic steps that involve noncovalent interactions without forming covalent intermediates with the enzyme molecules. These include:
   1. *Entropic effect*: Chemical catalysis in solution is slow because bringing together substrate and catalyst involves a considerable loss of entropy. The approximation and orientation of substrate within the confines of the enzyme-substrate complex in an enzymatic reaction circumvent the loss of translational or rotational entropy in the transition state. This advantage in entropy is compensated by the EA

binding energy. The rotational and translational entropies of the substrate are lost on formation of the EA complex, and not during the catalytic step.

2. *Strain and distortion*: An enzyme is able to accelerate a chemical reaction by having evolved an active site that is exquisitely complementary to the transition state. Although an active site must bind substrate, it is in the transition state that the binding interactions with the enzyme are maximized. To this objective, the substrate molecule is distorted to that of the transition state structure (Strynadka and James, 1991).
3. *General acid and base catalysis*: Protein enzymes often use a general base for partial removal of the proton from the attacking group, and a general acid for partial addition of a proton to the leaving group to stabilize the transition state (Figure 11.6). A crucial factor in the acid–base catalysis is that the catalytic

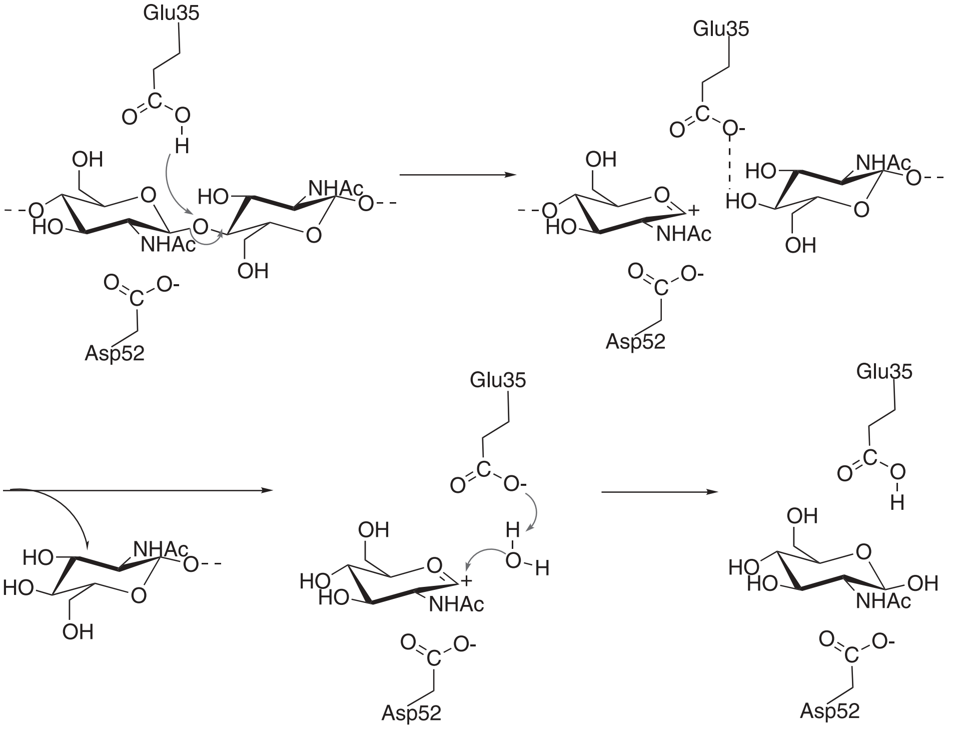
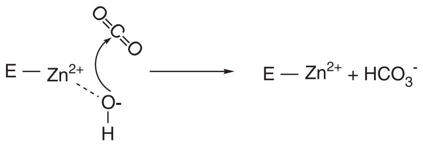


Figure 11.6 Mechanism of lysozyme catalysis
The lysozyme catalyzed hydrolysis (Phillips, 1967; Stryandka and James, 1991) demonstrates the involvement of noncovalent catalytic steps; N-Acetyl chitooligose is ushered into the active site cleft with an orientation along the six subsites (A to F subsites). The cleavage occurs between subsites D and E where the two catalytic residues, Glu35 and Asp52 are situated. The binding causes distortion of the sugar residue D to a half-chair-half-sofa conformation that of the transition state (Stryandka and James, 1991; Hadfield *et al.*, 1994). The carboxyl group of Glu35 is nestled within a predominantly nonpolar pocket and remains protonated, whereas the carboxyl group of Asp52 is surrounded by polar residues and is ionized. Glu35 acts as the general acid to protonate the glycosidic oxygen between the pyranose rings D and E (resulting in the deprotonation of Glu35). The cleavage of the glycosidic bond generates the oxocarbonium ion which is stabilized by Asp52 anion *via* ion-pair formation. The deprotonated carboxylate of Glu35, in turn, acts as the general base to abstract proton from a water molecule which hydroxylates/collapses with the oxocarbonium ion to complete the hydrolysis.

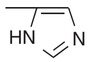
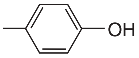
groups are in the correct ionization state under the reaction conditions. The most effective acid–base catalysts at pH 7 are those of pK_a ~7. This explains the widespread use of His (with an imidazole pK_a of 6–7) in enzyme catalysis.

4. *Electrostatic catalysis*: Electrostatic interactions are much stronger in nonpolar mediums than in water because of lower dielectric constants. Enzymes can stabilize ion pairs and other charge distribution effectively by virtue of the effective dielectrics (dielectric constant of 4 at the active site of the enzyme as compared to 80 for water) and oriented dipoles (from the hydrogen-bonded backbone groups of α helices) (Warshel, 1981). Furthermore, the charge groups at the enzyme active site may participate in the ion-pair formation, with the charge developed on the transition state (Figure 11.6).
5. *Metal ion catalysis*: Metal ions function in electrophilic catalysis stabilizing the negative charges that are formed. Metal-bound hydroxyl ions are potent nucleophiles (Strater *et al.*, 1996) that participate in reactions catalyzed by metalloenzymes (Christianson and Cox, 1999) and ribozymes (Cech and Bass, 1986).

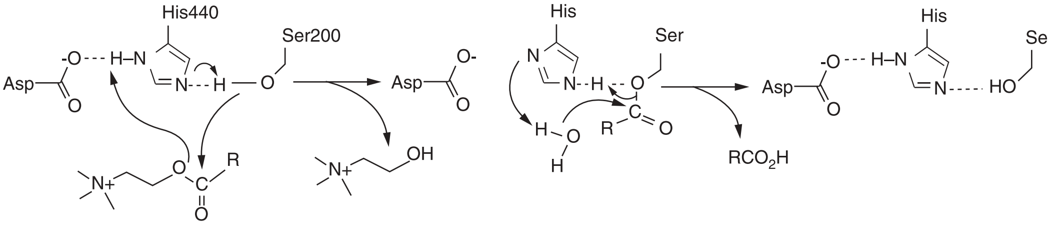


B. *Covalent catalysis*. In the catalytic steps that form covalent intermediates with the enzyme molecules, an initial attack commonly involves nucleophilic groups from the enzyme. Table 11.8 lists nucleophilic groups that are functional in enzyme catalyses. Covalent catalysis can be divided into three mechanistic classes:

TABLE 11.8 Nucleophilic groups in enzymes

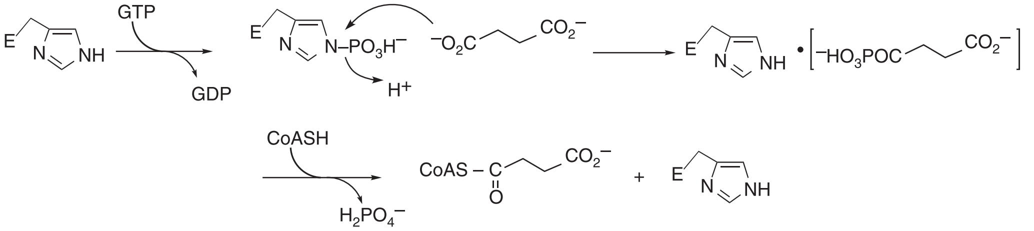
Nucleophile	Enzyme examples	Intermediate
Ser –OH	Serine protease	Acylenzyme
	Alkaline phosphatase, phosphoglucomutase	Phosphorylenzyme
Thr –OH	Amidase, proteasome	Acylenzyme
Cys –SH	Thiol protease, glyceraldehyde3-P dehdyrogenase	Acylenzyme
	Thymidylate synthetase	Uridylylenzyme
Asp $-\gamma CO_2^-$	K^+/Na^+–ATPase	Phosphorylenzyme
Glu $-\delta CO_2^-$	β-Retaining glucosidase, xylanases	Glycosylenzyme
Lys $-\varepsilon NH_2$	Acetoacetate decarboxylase, aldolase, pyridoxal enzymes	Schiff base
	DNA ligase	Adenylylenzyme
His 	Acid phosphokinase, histone phosphokinase, nucleoside diphosphokinase, phosphoglycerate mutase, succinyl-CoA synthetase	phosphorylenzyme
Tyr 	Glutamine synthetase, topoisomerase	Nucleotidylenzyme

1. The enzymatic nucleophile is similar in kind and reactivity to the ultimate solution acceptor. Examples of this class include the serine proteases and the alkaline phosphatase. The serine hydroxyl group is similar in chemical reactivity to the hydroxyl group of water, the final acceptor in these group transfer reactions (Fersht, 1985). For example, the active site Ser200 and His444 of cholinesterase are involved in a putative catalytic triad to effect acyl transfer (Taylor, 1991):

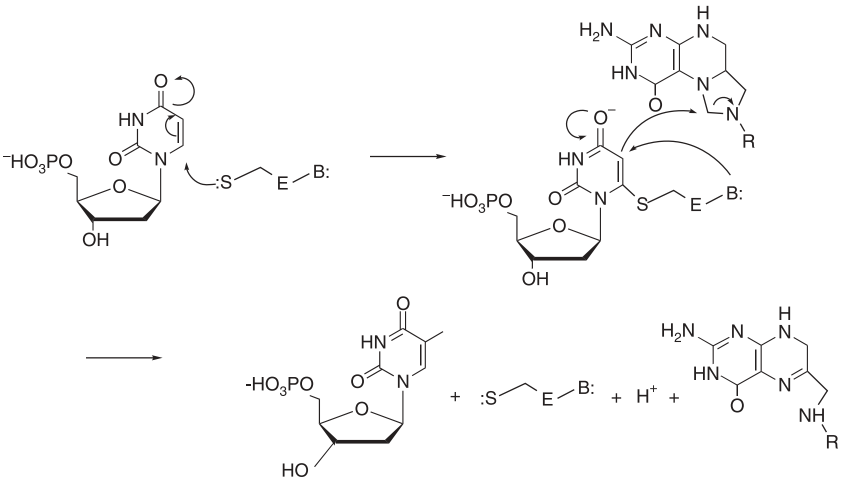


The advantage of this form of covalent catalysis presumably arises from the relative ease of positioning a nucleophile that is part of the protein itself and can be more easily handled in the interaction than a water molecule. Additionally, the serine ester intermediate is more reactive than the amide substrate, so that an attack by water is easier in the second step.

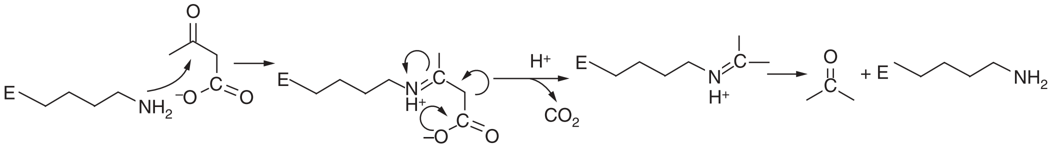
2. The enzymatic nucleophile is intrinsically more reactive than the ultimate acceptor. Several enzymes use the imdazole side chain of His as a transient acceptor in phosphoryl transfer reaction (Fersht, 1985). Amines react faster than alcohol with phosphoryl compounds, yet they form less stable adducts that can therefore readily allow rapid turnover.



3. Formation of a covalent adduct with an enzymatic nucleophile increases the reactivity of the substrate, which facilitates a reaction that is distinct from adduct formation. Examples of this include the thymidylate synthetase, in which the sulfhydryl of an active site Cys adds to C6 of the uracil. This breaks the aromaticity and adds electron density to increase the nucleophilicity of C5, thereby facilitating transfer of a methylene equivalent from tetrahydrofolate (Carreras and Santi, 1995).



Similarly, lysine residue form Schiff base intermediates with substrates, providing an electron sink in several enzymes, such as acetoacetate decarboxylase:



The stereospecificity is an important property of enzymes and the reaction stereochemistry of enzyme catalysis is a vital index of the organization of the enzyme active site and details of its reaction mechanism. The following rules, which determine stereochemical uniformity in enzymatic reactions, have been proposed (Hanson and Rose, 1975).

1. *Minimal number rule*: The enzyme reactions prefer to use one catalytic group whenever possible. This rule permits (a) one reactive group to catalyze multiple steps and (b) intrinsic proton recycling.
2. *Maximal separation rule*: A pair of catalytic groups attack a substrate molecule in an opposite rather than adjacent relationship such as the opposite distribution of acidic and basic groups encountered in many enzyme catalyses. This rule arises from:
   a) The active site commonly forms a cleft/crevice and critical catalytic groups are often situated on opposite walls of the cleft/crevice;
   b) A pair of catalytic groups in their reversible roles as acid and base is more effective on the opposite sides of an interposed substrate; and
   c) Anti-elimination catalyzed by opposite groups involves less motion.
3. *Minimal motion rule*: The observed stereochemistry of enzymatic reactions usually assumes that the proposed intermediates remain oriented through consecutive steps in a minimal motion relationship to the interacting groups. The selection for minimal motion can be explained:

a) Reversible enzymes must bind reactants, products and intermediates in a manner that approximates minimal motion.
b) Selection must occur for enzymes that effectively juxtapose reactive groups.
c) Selection occurs for processes in which the free energy of the transition state is minimized by stereoelectronic factors, such as maximal overlap of interacting orbitals.
d) Selection occurs for active sites in which the motions are partially frozen so that the relative loss of translational, rotational and internal entropy in going to a transition state is minimized.

Information regarding reaction mechanisms can be obtained by:

- *Kinetic studies*: sequences of substrate addition and product release, pH profiles of possible ionizable groups involved;
- *Chemical modifications and site-directed mutagenesis*: identification of catalytic and/or binding residues involved;
- *Physicochemical studies*: chemical transformations along the reaction paths based on known physicochemical principles;
- *Spectroscopic studies*: structural probe of the active site and characterization of reaction intermediates;
- *Chemical synthesis*: investigation of the transition state structures via synthetic transition-state analogues/inhibitors;
- *X-ray crystallography*: precise arrangement of atoms in the binding and catalysis;
- *Computational studies*: molecular modeling of possible and alternate reaction paths.

Kinetic studies of enzyme catalyses have been discussed. The special topics on X-ray crystallographic investigations of enzyme mechanisms have been reviewed (Lipcomb, 1983; Lolis and Petsko, 1990). We will now consider some of the common approaches employed to obtain information useful in the elucidation of enzyme reaction mechanisms.

11.4.2 Studies of enzyme mechanism: Active site

The kinetic studies on pH dependence of an enzymatic reaction yield a pH-rate profile, which may reveal the nature of ionizing groups involved in the catalysis. Six amino acid residues (His, Cys, Asp, Glu, Arg and Lys) account for more than 70% of all catalytic residues (Bartlett *et al.*, 2002). Enzymes utilize the limited set of these six side chains that form their catalytic toolkit (Gutteridge and Thonton, 2005). X-ray crystallographic studies of enzymes and their complexes (with substrates/products or their analogues and inhibitors) provide the most accurate information concerning the active site structures (Lipcomb, 1983; Lolis and Petsko, 1990). However, biochemists have employed other accessible techniques such as chemical modification (Sigman and Moosner, 1975; Ackers and Smith, 1985), site-directed mutagenesis (Smith, 1985; Gerlt, 1987; Profy and Schimmel, 1988; McPherson, 1991) and chemical manipulations (Dougherty, 2000; Hahn and Muir, 2005), in an attempt to characterize active sites of enzymes. Chemical manipulations such as expressed protein ligation, nonsense suppression and unnatural amino acid mutagenesis have been described in section 8.6.

11.4.2.1 Chemical modifications. Despite many obvious limitations, chemical modification had been the easiest approach to explore essential groups at the active site of an enzyme. However, chemical modification has been largely replaced by the site-directed mutagenesis and/or nonsense suppression approach, which is a much more reliable and accurate technique for the purpose. In view of the availability of a large volume of literature on the chemical modification studies that may provide useful information for preparation of modified enzymes with desirable industrial/pharmaceutical properties and design of activity-based probes in proteomic research, a brief presentation of chemical modifications of enzymes will be made.

Amino acid residues, except hydrocarbon chains, may provide nucleophilic sites (electron-rich centers) or electrophilic sites (electron-deficient centers) for chemical modifications. Electron-rich centers include sulfur nucleophiles (thiol of Cys and thioether of Thr), nitrogen nucleophiles (ε-amino of Lys, imidazole of His and Guanidyl of Arg), oxygen nucleophiles (phenolic of Tyr, carboxyl of Asp and Glu and hydroxyl of Ser and Thr), and carbon nucleophile (α-position of indole ring of Trp), with an increasing nucleophilicity in that order. They provide nucleophilic sites for alkylation (nucleophilic substitution), acylation, addition and oxidation at pH near or above their pK values. Electron-deficient centers include ammonium cation of Lys, guanidium cation of Arg and imidazolium cation of His. They provide electrophilic sites for metalation and reduction at pH near or below their pK values.

The chemical properties of a protein functional group are strongly influenced by its local environment, such as polarity of the microenvironment, hydrogen bonding effect, field/electrostatic effect, as well as steric and matrix effects. These factors contribute to the selectivity and specificity of chemical modifications. Several strategies can be adapted to enhance the selectivity and specificity:

1. the use of amino acid-selective reagents;
2. the exploitation of enhanced or differential reactivity of amino acid residues;
3. he application of differential or sequential protection; and
4. the exploitation of protein specificity such as affinity labeling (Wold, 1977) and suicide reagent (Abeles and Maycock, 1976).

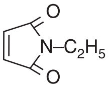
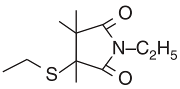
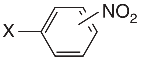
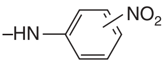
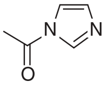
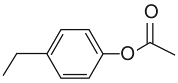
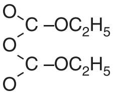
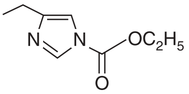
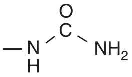
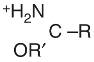
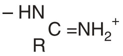
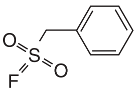
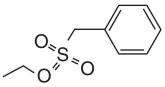
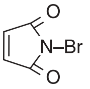
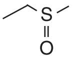
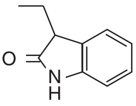
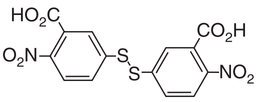
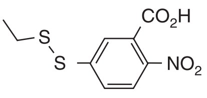
Table 11.9 lists some of chemical modifications commonly used to explore essential groups of enzymes (Means and Feeney, 1971).

11.4.2.2 Site-directed mutagenesis. An *in vitro* mutagenesis, which puts mutations precisely where they are needed, is termed site-directed mutagenesis (Smith, 1985; Shaw, 1987). The technique allows researchers to execute any modification at any position desired in cloned DNA. It provides a powerful tool for the analysis of active sites of enzymes. It is generally carried out via oligonucleotide-directed or cassette approaches. In the site-directed mutagenesis, preselected codon change(s) are introduced into the coding region of a cloned gene such that the expressed enzyme contains the corresponding amino acid replacement. This method of perturbing the active sites of enzymes overcomes many of the limitations associated with chemical modifications. Obvious advantages of the site-directed mutagenesis are:

1. The amino acid substitution introduced are specific and quantitative.
2. A wide variety of amino acid replacements can be performed.
3. Protein molecules with subtle structural and functional changes can be prepared.

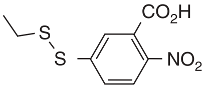
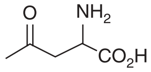
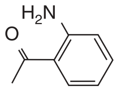
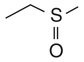
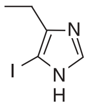
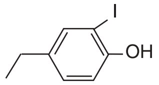
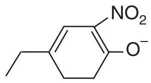
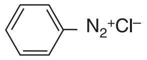
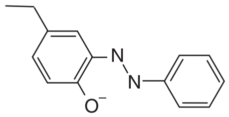
Tables 11.10 and 11.11 illustrate studies of enzyme active sites by the site-directed mutagenesis.

TABLE 11.9 Commonly used chemical modifications of amino acid functional groups

Reaction	Chemical reagent	Residue modified	Product of main residue modified	Remark
Alkylation	$ICH_2CO_2^-$	C ≫ H > M > K	$–S\text{-}CH_2CO_2^-$	Mainly C at pH ≥ 7, however H (pH > 5.5) and K (pH > 8.5) may react
		C, K		C at pH > 5 and K at pH > 7
		K > C > Y > H		Reactivity for X = F > Cl≈Br > SO_3H
Acylation	RCHO + $NaBH_4$	K	$–CH_2R$	Reductive alkylation
		Y, K		pH 7.5
		H, K		pH 4–7
	HN=C=O	K, C, Y, D/E, H H		Mainly K at pH ≥ 7, pH for C (6–8), D/E (~5), H (~8)
		K		pH ≥ 8.5
		K, S		Both tosyl chloride and tosyl fluoride can be used
Esterification	$R'N{=}C{=}NHR^+$	D, E		
	+ ROH		$–CO_2R$	Ester with ROH
Amidation	+ RNH_2		–NHR	Amide with RNH_2
Oxidation	H_2O_2	C	$–SO_3H$	
		M		
		W		
		C		With adjacent Cs, disulfide, -S–S- may be formed

(continued)

TABLE 11.9 ***continued***

Reaction	Chemical reagent	Residue modified	Product of main residue modified	Remark
Photooxidation	Rose bengal (anionic) or methylene blue (cationic) as sensitizer	H		Neutral pH (~7)
		W	 	Low pH (<4)
		M		Low pH (<4)
Electrophilic Substitution	I_3^-	C	SO_3^-	Final oxidized product.
		H		
		Y		
	$C(NO_2)_4$	Y		
		Y		pH ~9 at 0°C

Note: Amino acid residues, which are modified, are given in one-letter abbreviations. The modified products (attached to polypeptide chain) are shown with the modified groups for the major residues affected.

11.4.2.3 Spectroscopic probe. Various spectroscopic techniques have been employed to probe the active site and therefore to delineate the reaction mechanism of an enzymatic reaction. X-ray crystallography provides the ultimate means of investigating the active sites and reaction mechanisms of enzymes. Exploration of the active sites and their reaction investigated by other spectroscopic tools can sometimes produce fruitful results.

The UV/visible spectroscopic probe consists of protein residues (e.g. Trp, Tyr), cofactors/coenzyes (e.g. pyridoxal and flavin coenzymes) and substrates or their analogs as target chromophores. For example, pyridoxal-5′-phosphate (PLP) of pyridoxal enzymes undergo different stages of transformations that can be monitored spectroscopically (Metzler, 1979) as shown in Figure 11.7.

Infrared spectroscopy has only limited use in probing the active sites and reaction intermediates of enzymes because of the interference from water. For example, the carbonyl band ($\nu_{C=O}$) for free dihydroxyacetone phosphate is centered around 1733 cm^{-1}.

TABLE 11.10 Studies on catalytic residues of lysozymes by site-directed mutagenesis

Lysozyme	Catalytic residue	Mutant	Observation (% activity of the wild type)	Possible explanation
Hen's egg white	Asp52	D52N	No measurable activity	Both E35 and D52 are required for the catalytic activity. Large reduction in activity without anionic D52 to form ion-pair. NAG_4 binds to the A–D subsites of D52S with the sugar in site D having sofa conformation and α configuration.
	Asp52	D52S	Less than 1%, Pyranose ring in site D of D52S-NAG_4 has a sofa conformation.	
	Asp52	D52A/S/T/C	Lytic activity decreases (% of wild) to 1.5, 0.71, 0.91, 0.62	
	Glu35	E35Q	~5%	
T4	Asp20	D20C/E	Nearly active as the wild type for D20C but greatly reduced activity (1.2% of wild) for D20E	The residue at 20 (D20/C20) has a significant nucleophilic function. Critical distance between the catalytic residue and scissile bond.
Pseumococcus	Asp9	D9E/N/K	Catalytic activity (% of wild) reduced to 1.7, 2.2 and 0	The carboxyl groups are essential. Replacement with basic K abolishes the activity. The residue D9 is the critical catalytic residue. Furthermore the distance between D9 and the scissile bond is critical.
	Glu36	E36D/Q/K	Catalytic activity (% of wild) reduced to 37, 67 and 0.003	

Notes: 1. Mutants are expressed (one-letter abbreviation for amino acids) as: wild-#position-mutant, i.e. E35Q denotes the replacement of Glu35 for Gln in the mutant; D52A/S/T/C denotes mutants with D52A, D52S, D52T and D52C replacements.
2. References for hen's egg white lysozyme, Malcolm *et al*. (1989), Hadfield *et al*. (1994) and Hashimoto *et al*. (1996); T4 lysozyme, Hardy and Poteete (1991); *Pseumococcus* lysozyme, Sanz *et al*. (1992).

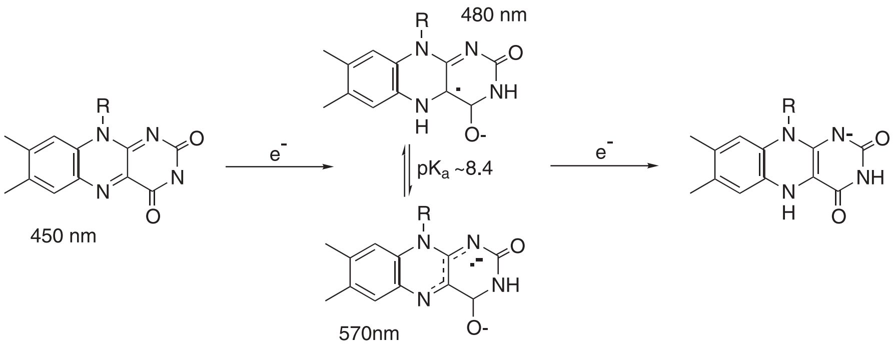
However, two carbonyl bands, $\nu_{C=O} = 1713\,cm^{-1}$ and $1732\,cm^{-1}$ (with an intensity ratio of about 3 : 1), corresponding to the bound dihydroxyacetone phosphate to triosephosphate isomerase, are observed. A shift of the major band by $19\,cm^{-1}$ provides plausible evidence of enzyme-induced distortion of the substrate. This is probably attributable to an enzyme electrophile that polarizes the carbonyl group of dihydroxyacetone phosphate and thereby promotes catalysis. Resonance Raman spectroscopy has been used to obtain the vibrational spectra of the acyl carbonyl groups of acyl-subtilisins. As the deacylation rate increases, the carbonyl stretching band ($\nu_{C=O}$) of the acyl-subtilisin is observed to shift to a lower frequency, indicating an increase in the single bond character of the reactive acyl carbonyl group. It is estimated that $r_{C=O}$ increases by 0.015 Å as the deacylation rate increases 500-fold through the series of acyl enzymes.

Nuclear magnetic resonance (NMR) spectroscopy is a choice in solution for delineating the active sites and enzyme complexes, including the specific interactions between substrates/products with specific protein residues and the disposition of substrates/products within the active site (Cohn and Reed, 1982). Dynamic phenomena, such as interconverting protein conformations or substrate and product complexes, can lead to spectral effects that are interpretable in terms of reaction mechanisms. For example, the charge

TABLE 11.11 Studies on enzyme active sites by site-directed mutagenesis

Enzyme	Critical residue	Mutant	Observation	Possible explanation
Rat trypsin	Asp102	D102N	Rel. activity (D102N/wild): k_{cat} = 0.0002, Km = 2 at pH7.0	Isosteric replacement of Asp102 of the triad (D102, H57, S195) reduces the catalytic activity.
B. amyloliquefaciens subtilisin BPN′	Asp32 His64 Ser221	D32A H64A S221A	Reduction of k_{cat} by factors of 3×10^4, 2×10^6 and 2×10^6 respectively	Importance of catalytic triad, D32, H64 and S221 in the subtilisin catalysis.
Yeast triosephosphate isomerase	His95	H95Q	Reduction of k_{cat} (GAP) by 1/380 and k_{cat} (DHAP) by 1/140.	Substitution of electrophilic H95 impairs the ability to stabilize the enediol intermediate.
E. coli glyceraldehyde-3-phosphate dehydrogenase	H176	H176N	k_{cat} for H176N is ~1/60 for oxid. Phosphorylation and 1/45 for the reverse reaction	H176 enhances the nucleophilicity of C149 and acts as a hydrogen donor to facilitate the formation of tetrahedral intermediate.
Human lipoamide dehyddrogenase	 His452 Glu457	Wild H452Q E457Q	Ping pong bi bi mechanism Ping pong bi bi mechanism Order bi bi mechanism	Demonstration that an essential residue may affect the kinetic mechanism of enzyme reaction.
B. stereothermophilus lactate dehydrogenase	Asp53	D53S	Increase NADPH activity by 20 times	The anionic Asp is responsible for the coenzyme specificity of NAD^+ dependent dehydrogenasess. The replacement of Asp abolishes the discrimination between NAD(H) and its 2′-phosphate derivative, NADP(H).
Yeast alcohol dehydrogenase	Asp201	D201G	Almost equal utilization of NAD^+ and $NADP^+$	
E. coli glutathione reductase	 Arg198 Arg204	Wild R198M R204L	k_{cat}/K_m (NADPH & NADH) = 740.0 & 0.34 for wild, 8.5 & 0.96 for R198M, 12.8 & 0.17 for R204L	The importance of conserved Arg198 and Arg204 in NADP(H)-flavo-oxidoreductases in binding of the 2′-phosphate group of NADPH.
E. coli alkaline phosphatase	Asp153	D153H	Mg^{2+} is replaced by Zn^{2+}	Conversion of the octahedral Mg^{2+}-binding site to a tetrahedral Zn^{2+}-binding site in which an imidazole N of His153 serves as a ligand.
B. stereothermophilus tyrosyl-tRNA synthetase	Thr40 His45	T40A/G H45A/G/N	Increase in the binding energy level of the transition state by ~20 kJ/mol	Thr40 and H45 bind Tyr and ATP in the transition state of the reaction.

Note: References used are: Craik *et al.* (1987) for rat trypsin; Carter and Wells (1988) for subtilisin; Nickbarg *et al.* (1988) for yeast triosephosphate isomerase (GAP for glyceraldehyde-3-phosphate and DHAP for dihydroxyacetone phosphate); Soukri *et al.* (1989) for *E. coli* glyceraldehyde-3-phosphate dehydrogenase; Kim and Patel (1992) for human lipoamide dehydrogenase; Fan *et al.* (1991) for yeast alcohol dehydrogenase; Scrutton *et al.* (1990) for *E. coli* glutathione reductase; Murphy *et al.* (1993) for *E. coli* alkaline phosphatase; Leatherbarrow *et al.* (1985) for tyrosyl-tRNA synthetase.


480 nm
e-
pKa ~8.4
e-
450 nm
570nm

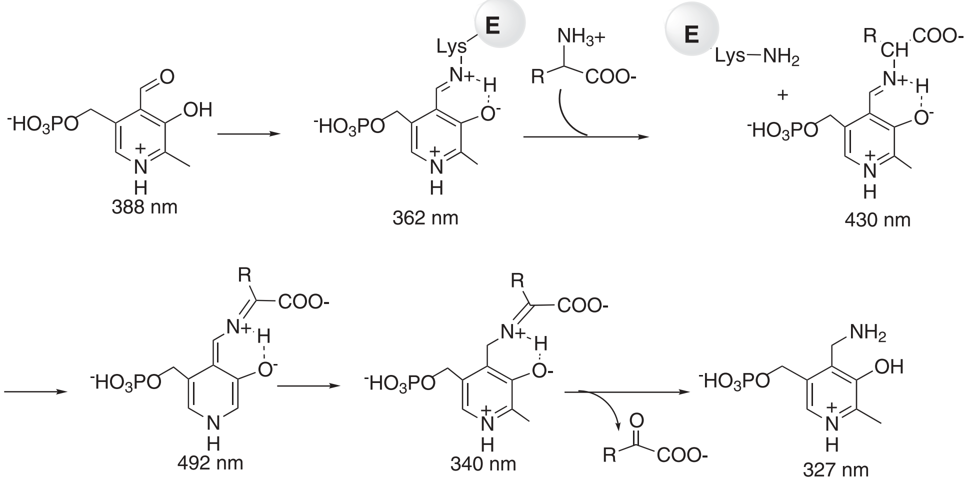
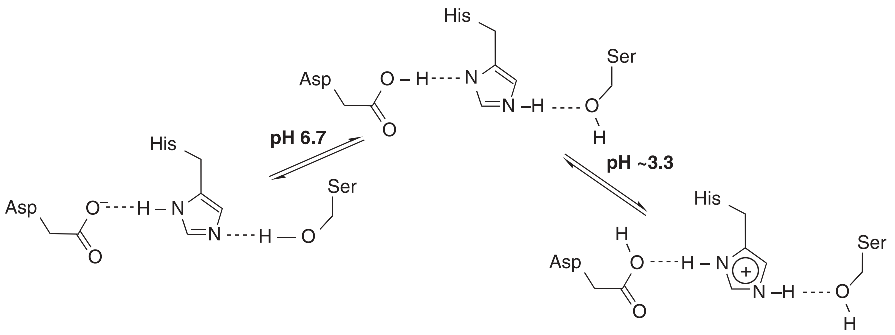

E
Lys
E
Lys—NH2
388 nm
362 nm
430 nm
492 nm
340 nm
327 nm

Figure 11.7 Spectral changes accompanying pyridoxal enzyme reaction
The spectral changes accompanying transamination catalyzed by α-amino acid aminotransferase and the plausible intermediates associated with absorption bands (λ_{max}) are shown.

relay system in serine proteases can be explicitly demonstrated by NMR spectroscopic approach. α-Lytic protease (serine protease) from *Myxobacter* contains only a single His, which can be selectively enriched with 2-[13C]-His. ^{13}C-NMR (CMR) is carried out on [^{13}C]-α-lytic protease to demonstrate over the pH range; the catalytic triad as shown:


His
Ser
Asp
pH 6.7
pH ~3.3
His
Ser
Asp
His
Ser
Asp

The His residue play two roles:

1. It provides insulation between water and the buried carboxylate anion of Asp, thus ensuring the carboxylate group a hydrophobic environment with an elevated pK value.
2. It provides a relay for net transfer of a proton from the Ser-hydroxyl to the Asp-carboxylate anion.

Phosphorus magnetic resonance spectroscopy (^{31}PMR) is a useful tool for monitoring the binding and dynamics of P-containing coenzymes and substrates or products during enzyme catalysis. For example, the bound 2′-phosphate group of NADP/NADPH ($\delta = 6$ ppm) in the NADP(H)-*iso*citrate dehydrogenase complex is shown to be in the dianion form, interacting with a positively charged residue or distortion in its P—O—P linkages, due to the downfield shift of 1.8 ppm. The phosphate group that is transferred in the catalytic step between G1P and G6P is a dianion involving Ser116 of phosphomutase. Relaxation NMR techniques with paramagnetic probes (Mildvan and Cohn, 1970) have provided important information concerning coordination structures of metal ions in metalloenzymes.

11.4.3 Studies of enzyme mechanism: Transition state

All chemical transformations pass through an unstable structure called the transition state, which is poised between the chemical structures of the substrates and products, whereas the reaction intermediate is a transient stable structure that is formed along the reaction path to or from the transition state. The transition state structure of an enzymatic reaction cannot be observed directly, whereas the reaction intermediate can be observed or can often be isolated/trapped for characterization.

Transition state theory for enzyme-catalyzed reactions has its origins in the theory of rate processes (Glasstone *et al.*, 1941) that molecules possessing the highest energy structure found along the reaction path can be considered to constitute the transition state. This activated state is defined by an unstable vibrational coordinate that decomposes to reactants and products. Enzymes can act by binding tightly to and specifically stabilizing the transition state molecules. Those residues that are involved in catalysis of an individual step preferentially bind one of the transition states in the reaction. An advantage is gained only when the energy barrier of the rate-limiting step is lowered. The active site of an enzyme is probably most complementary to the highest transition state along the reaction coordinate. Transition states are unstable and reactive peaks, whilst reaction intermediates are transiently stable, trough in the energy profile along the reaction coordinate. Two explanations regarding the ability of the active site to bind multiple transition states as well as the substrate:

1. For situations where the active site is relatively rigid, all chemical species along the reaction coordinate must bear a close resemblance to one another, so the enzyme has no difficulty in binding all of them. The strain applied by the active site due to its complementarity to one of the transition states forces the substrate into a conformation toward that transition state where all favorable interactions are finally maximized.
2. Some enzymes are known to undergo large conformation changes during catalysis. It is likely that the flexibility allows the active site to adapt to each species along the reaction coordinate.

The transition state structures of enzyme-catalyzed reactions provide fundamental information about the interactions used to catalyze biological reactions. All chemical transformations pass through the transition states, which are proposed to have lifetimes near 10^{-13} s (equivalent to the time for a single bond vibration). Enzymes typically increase the rate over the noncatalyzed reaction by factors of 10^{10}–10^{15}. The enzyme-substrate complexes often have dissociation constants in the range of 10^{-3}–10^{-6} M and it has been proposed that transition state complexes are bound with dissociation constants in the range of 10^{-14}–10^{-23} M (Schramm, 1998).

No physical or spectroscopic methods are available for the direct observation of the transition state structure for enzymatic reactions. Yet transition state structure is central to understanding catalysis, because enzymes function by lowering activation energy. Experimental approaches to investigate the transition state structure of enzymatic reactions include:

1. *Chemical precedent* (Walsh, 1979; Jencks, 1987): Behavior of model chemistry in solution for the reaction of interest provides a transition state benchmark for comparison with the transition state structure imposed by the enzyme. This information leads to a design to trap transition state or to synthesize transition state inhibitors.
2. *Kinetic isotope effects* (Suhnel and Schowen, 1991): Kinetic isotope effects compare the enzymatic reaction rates of isotopically labeled and unlabeled substrates. Isotopically labeled molecules have molecular energy different from that of unlabeled molecules and thus require a different amount of energy to reach the transition state, where the molecular energies may also be perturbed by the isotope. If the bonding environment for the labeled atom is less restricted in the transition state than in the reactant, the isotope effect will be normal (the heavy isotope substrate reacting more slowly than the unlabeled substrate). Otherwise, if the bonding environment for the labeled atom at the transition state is more restricted than in the reactant, the substrate with the higher mass isotope will react more rapidly (an inverse kinetic isotope effect). Isotope effects also provide quantitative information because the magnitude of the isotope effect indicates the extent of bond change. By measuring kinetic isotope effects at every position in a substrate molecule that might be expected to be perturbed at the transition state, a unique description of the transition state can be deduced if intrinsic isotope effects are being measured (Northrop, 1981).
3. *Transition state inhibitors* (Wolfenden, 1972; Morrison and Walsh, 1988): Inhibitors that bind tightly to the enzymes and resemble the expected transitions states have been used to predict structural features of the transition states. The design of transition state inhibitors is aided by the calculation of electrostatic potential surfaces of expected transition states (Schramm, 1998).
4. *Inference from the trapped intermediates*: Several enzymes, particularly hydrolases, form covalent enzyme intermediates such as acyl enzyme intermediates (Bell and Koshland, 1971), which are proximal to the transition states in the energy profile along the reaction coordinate. Since the reaction intermediates, in some cases, can be trapped/isolated for characterization, they represent informative models for the transition states of enzymatic reactions.

11.4.4 Structure-activity relationship

Studies on the structure-activity relationship of a chemical reaction often provide valuable information regarding its chemical mechanism. Quantitative structure-activity relationship

(QSAR) represents an attempt to correlate structural or property descriptors of compounds with reactivities (Topliss, 1983; Hansch and Hoekman, 1995). An explanation of the structural effect on the equilibria or rates of chemical reactions often involve some kind of comparison with a suitable model or set of models. The quantities compared are thermodynamic, usually free energies, enthalpies or entropies, but the simple relationships found among such quantities are often not part of the formal structure of thermodynamics, hence they are referred to extrathermodynamic relationships. Useful extrathermodynamic relationships are usually simple in form. The mathematical simplicity of many extrathermodynamic relationships results from the tendency of such quantities to be additive functions of molecular structure. The molar values of the property under comparison are assumed to be approximately additive functions of independent contributions assignable to substructures of the molecules.

The best known of the extrathermodynamic equation based on linear free energy relationship on the reaction rate or equilibrium of the aromatic system to the structure of the reagent, is the Hammett equation:

$$\log(k_x/k_0) \text{ or } \log(K_x/K_0) = \rho\sigma$$

where k_0, k_x are rate constants and K_0, K_x are equilibrium constants for the model compound (x = H) and compound with substituent x respectively. The Hammett equation describes the effect of a *meta*- or *para*-substituent on the rate or equilibrium constant of an aromatic side-chain reaction. It is based on the fact that, as the substituent is varied, the logarithms of the rate or equilibrium constants for a large number of aromatic side-chain reactions are linearly related to one another under the same experimental conditions. σ is the electronic constant that exerts resonance and/or inductive effects for the *para*-substituent (σ_p) and/or *meta*-substituent (σ_m). These values for various substituents are given in the Table 11.12. ρ is reaction constant, which is a function of the reaction conditions. When two or more substituents are introduced simultaneously into *meta* or *para* positions of an aromatic compound, it is usually found that their combined effect can be represented by the sum of their individual sigma values.

For the substituent constant, a positive value denotes electron-withdrawing and a negative value indicates electron-donating for the substituent. It therefore appears for the substituent effect in series involving only a single interaction mechanism that a positive value of ρ indicates a reaction in which negative charge (or partial negative charge) is developed on the side chain, whereas a negative value for ρ corresponds to the development of a positive charge (or partial positive charge) on the side chain. When the mechanism of a reaction changes because of the presence of certain substituents or when the measured rate constants is actually a composite quantity depending on the rate and equilibrium constants of several reaction steps, curvature in the $\rho\sigma$ relationship can occur. A change in mechanism always causes the curve ($\rho\sigma$ plot) to be concave up, while a change in rate-limiting step with otherwise constant mechanism can cause the curve to be concave down.

An analogous relationship (Hammett equation) is ascribed to a large number of aliphatic reaction rates:

$$\log(k_x/k_0) = \rho^*\sigma^*$$

in which ρ^* is an empirical parameter dependent on the nature of the reaction and the reaction conditions. Formally the $\rho^*\sigma^*$ equation is a correlation of free energies by the method of linear combination. For some reaction series it is possible to correlate derivations from the $\rho^*\sigma^*$ equation apparently due to steric effect by establishing a set of steric

TABLE 11.12 Substituent constants of some common substituents

Substituent	σ_p	σ_m	σ^*	E_s	π
H	0.00	0.00	0.49	1.24	0.00
Br	0.23	0.39	2.80	0.08	0.86
Cl	0.23	0.37	2.68	0.27	0.71
F	0.06	0.34	3.08	0.78	0.14
I	0.18	0.35	2.38	–0.16	1.12
NO_2	0.78	0.71		–1.28	–0.28
OH	–0.37	0.12	1.55	0.69	–0.67
SH	0.15	0.25		0.17	0.39
NH_2	–0.66	–0.16		0.63	–1.23
CN	0.66	0.56	3.64	0.73	–0.57
CHO	0.42	0.35			–0.65
CO_2H	0.45	0.37	2.94		–0.32
$CONH_2$	0.36	0.28			–1.49
CH_3	–0.17	–0.07	0.00	0.00	0.56
OCH_3	–0.27	0.12	1.81	0.69	–0.02
CH_2OH	0.00	0.00	0.56	0.03	–1.03
SCH_3	0.00	0.15	1.47	0.17	0.61
$NHCH_3$	–0.84	–0.30			–0.47
$COCH_3$	0.50	0.38	1.65		–0.55
$COOCH_3$	0.45	0.37	2.00		–0.01
$NHCOCH_3$	0.00	0.21			–0.97
CH_2CH_3	–0.15	–0.07	–0.10	–0.07	1.02
$CH_2CH_2CH_3$	–0.13	–0.07	–0.12	–0.36	1.55
$CH(CH_3)_2$	–0.15	–0.07	–0.19	–0.47	1.53
$N(CH_3)_3^+$	0.82	0.88			–5.96
$C(CH_3)_3$	–0.20	–0.10	–0.30	–1.54	1.98
$(CH_2)_3CH_3$	–0.16	–0.08	–0.13	–0.39	
C_6H_5	–0.01	0.06	0.60		1.96
OC_6H_5	–0.03	0.25	2.24		2.08
$COO\ C_6H_5$	0.13	0.21			1.46

Note: The QSAR analysis may be performed online at QSAR server (http://mmlin1.pha.unc.edu/~jin/QSAR/).

substituent constants, E_s. For reactions in which the steric effects of substituents are involved, the following Taft equation is applicable:

$$\log(k_x/k_0) = \rho^*\sigma^* + sE_s$$

where s is the steric susceptibility constant with a positive s value indicating a retardation of the rate process by steric hindrance. The Taft equation is another correlation by linear combination in which the relative importance of two interaction mechanisms (ρ^* and s) is involved. A useful approach for predicting the dependence on the same steric interaction mechanisms for two systems is the principle of isosterism. It states that the steric mechanisms should be identical in two structures of the same size and shape, regardless of the particular elements from which they are formed.

There might be a quantitative relationship between biological reactivity and any number of chemical properties in a series of biochemicals. According to the extrathermodynamic relationship, it is assumed that the relative biological effect of members of a set of analogs is due to the effect of the variable substituents on the chemical and physical properties of the molecule. For biological activity, an important property of a molecule is its partition coefficient, P, which can be defined as the equilibrium constant for the distribution of the biomolecule (B) between hydrophobic (Ψ) and aqueous (ω) phases, i.e.:

$$P = [B]_{\Psi}/[B]_{\omega}.$$

Since the extrathermodynamic method investigates correlation between thermodynamic properties, the appropriate physical property is the relative free energy of the above equilibrium. Hence $\log P$ is used to express this physical property, known as hydrophobicity (or lipophilicity).

Hydrophobicity is measured as the relative affinity of a molecule for a nonpolar phase versus that for water. Octanol is chosen as a solvent for the measurement of partition coefficients because its hydroxyl group will make it similar to a biological membrane, which also has hydrogen bond donating and accepting properties. The logarithm of the octanol-water partition coefficient is used to define the hydrophobic parameter, π, i.e.:

$$P = [B]_{\text{octanol}}/[B]_{\text{water}}$$
$$\pi_x = \log P_x - \log P_H$$

in which P_x and P_H are octanol-water partition coefficients for a compound with a substituent X and the parent (model) compound respectively. Thus the hydrophobic effect on the biological reactivity ($1/C$ or $1/K_m$, $1/K_i$) of a series of compounds with substituents can be expressed by the Hansch equation (Fujita *et al.*, 1964) as

$$\log(1/C) \text{ or } \log(1/K_m) = \phi\pi_x + d$$

The parameter, ϕ represents the hydrophobic contribution to biological activities, $1/C$, in which C is the molar concentration of the compound that produces a standard biological response. Hydrophobic facilitation of the biological activity is characterized by a positive ϕ value, while hydrophilic facilitation is associated with a negative ϕ value.

Since a substituent is produced in biological systems, at least three major changes occur in the chemical properties of a molecule, i.e. electronic, steric and hydrophobic. The substituent effect on the biological properties of a molecule can result from the change of some or all of these chemical properties. The QSAR of their effects on biological activity are described by the combined equation, which has been applied to analyze the catalytic constants and rates in mechanistic studies of enzymatic reaction (Hansch and Klein, 1991; Tsai *et al.*, 1969):

$$\log 1/K_a \text{ or } \log k_{cat} = \rho\sigma + sE_s + \phi\pi + d$$

In addition to the electronic (σ), steric (E_s) and hydrophobic (π) descriptors (Table 11.12), other physical and chemical descriptors such as electronic descriptors (e.g. ionization constants pK_a, NMR chemical shifts δ) steric descriptors (e.g. molar volume MV, van der Waals volume V_w, molar refractivity MR, parachor P_r), hydrophobic descriptors (e.g. distribution coefficient $\log D$, solubilty parameter $\log S$), property descriptors (e.g. T_m of DNA) and theoretical parameters (e.g. atomic charge q^{σ} or q^{π}, electrostatic potential $V(r)$) are also employed in the QSAR studies of chemical and biological systems.

11.4.5 X-ray crystallographic studies and refinement

Chemical modification and mutagenesis experiments can identify amino acids in the active site that participate in catalysis. The precise arrangement of atoms in an active site is determined by X-ray crystallography (Lipcomb, 1983). Structures of enzymes complexed to catalytically relevant ligands can reveal the interactions that are responsible for catalysis. Although transition states have a short lifetime ($\sim 10^{-13}$ s, therefore are unstable and not amendable to isolation/characterization), molecules that mimic the transition state of an enzyme catalyzed reaction should bind tightly to the enzyme and could be used as a model of the transition state. The X-ray structures of transition-state analogues bound to enzymes may provide structural basis for formulating catalysis. A complete dynamic picture of the reaction mechanism, including any conformational changes, should emerge (Lolis and Petsko, 1990) when combined with structures of the uncomplexed enzyme and of the enzyme complexed to substrates/products and intermediates. Structural bases of catalyses derived from the X-ray investigations for several representative enzymes will be described in the following subsection.

Molecular modeling represents a theoretical approach to the refinement of enzyme reaction mechanisms (Kollman, 1985; Brünger and Karplus, 1991). The method can be divided into two classes:

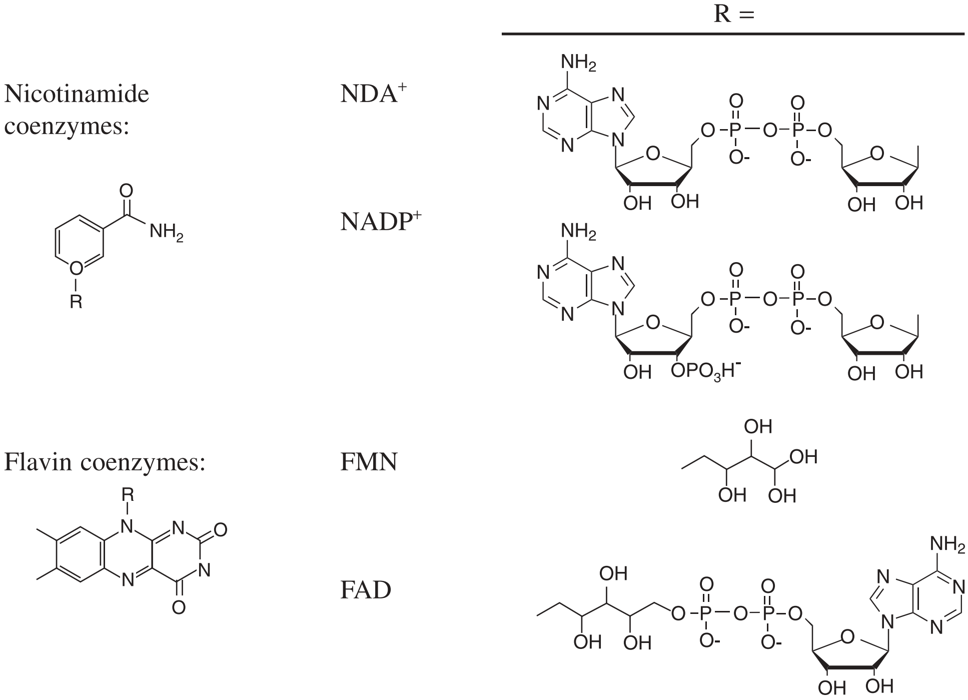
1. Model building, in which molecular structure is represented by experimental data as input and this structure is manipulated with use of stereochemical rules. This includes computer graphics and distance geometry refinement.
2. Energy functions include *ab initio*, semiempirical quantum and molecular mechanical calculations to effect energy minimization and molecular dynamic simulations. The application of molecular mechanical calculation has been discussed in Chapter 9.

Computer graphics in combination with numerical calculations have allowed us to understand and visualize enzymatic reactions at a molecular detail. While results of computer-based theoretical approaches may require further improvement and verification, they either reinforce the experimental views or suggest directions for new experiments. For example, mechanical calculation and dynamic simulation of papain (cystein protesase) catalysis suggest that the nucleophilic attack by the thiolate anion of Cys25 gives rise to a protonated tetrahedral transition state (Arad *et al.*, 1990) rather than an anionic tetrahedral transition state analogous to the serine protease. Molecular mechanical calculation implies that an alternative mechanism (i.e. glycosyl enzyme intermediate), in addition to the oxocarbonium mechanism, is also possible for the lysozyme catalyzed hydrolysis of synthetic substrates (Tsai, 1997).

11.4.6 Case studies of enzyme mechanisms

A large number of enzyme mechanisms, including the enzyme structures, kinetic behavior and the overall chemical pathways of the reactions are known (Sinnott, 1998). Herewith, examples from each enzyme classes are given:

11.4.6.1 Oxidoreductase: dehydrogenase and flavoenzymes. Nicotinamide adenine dinucleotide (NAD^+)/nicotinamide adenine dinucleotide phosphate ($NADP^+$) and flavin mononucleotide (FMN)/flavin adenine dinucleotide (FAD) are the two major types of redox coenzymes.



Dehydrogenases (DHs) usually show a preference for either NAD$^+$ or NADP$^+$. The critical amino acid residue, Arg, which interacts with the 2′-phosphate group of the adenosine moiety of NADP$^+$ for the NADP$^+$-specific dehydrogenases (e.g. G6PDH, 6P-gluconate DH) is missing or displaced by an Asp/Glu in the NAD$^+$-preferred dehydrogenases (e.g. alcohol DH, lactate DH). In some cases, both NAD$^+$ and NADP$^+$ preferred enzymes are found (e.g. *iso*citrate DHs, glycerol-3P DHs, glutamate DHs). Oxidation of the reduced substrates in the NAD$^+$/NADP$^+$-dependent enzymatic reactions involves the removal of two hydrogen atoms. One is transferred directly as a hydride ion (H$^-$) to the 4 position of the nicotinamide ring and the other is released as a proton (Figure 11.8). The hydride transfer is prochiral stereospecific (You, 1982). The stereospecificity of hydride transfer to the nicotinamide coenzyme is proR for some (e.g. alcohol DH, aldehyde DH, L-lactateDH, malate DH, *iso*citrate DH) but proS for others (e.g. Glucose DH, G6PDH, glutamate DH, lipoamide DH).

Flavoenzymes catalyze a large number of diverse redox interconversions reflecting the chemical versatility of the tricyclic isoalloxazine ring system of the flavin coenzymes, which act as a two-electron/one-electron switch as well as a molecular oxygen activator (Ghisla and Massey, 1989). Therefore flavin coenzymes serve:

- as a step-wise transforming redox switches between obligate two-electron donors and obligate one-electron acceptors; and
- as cofactors for net two electron reduction of O_2 to H_2O_2 or activation/cleavage of O_2 in mono-/di-oxygenation, due to their reactivity with molecular oxygen.

These redox reactions may be placed in one of three categories:

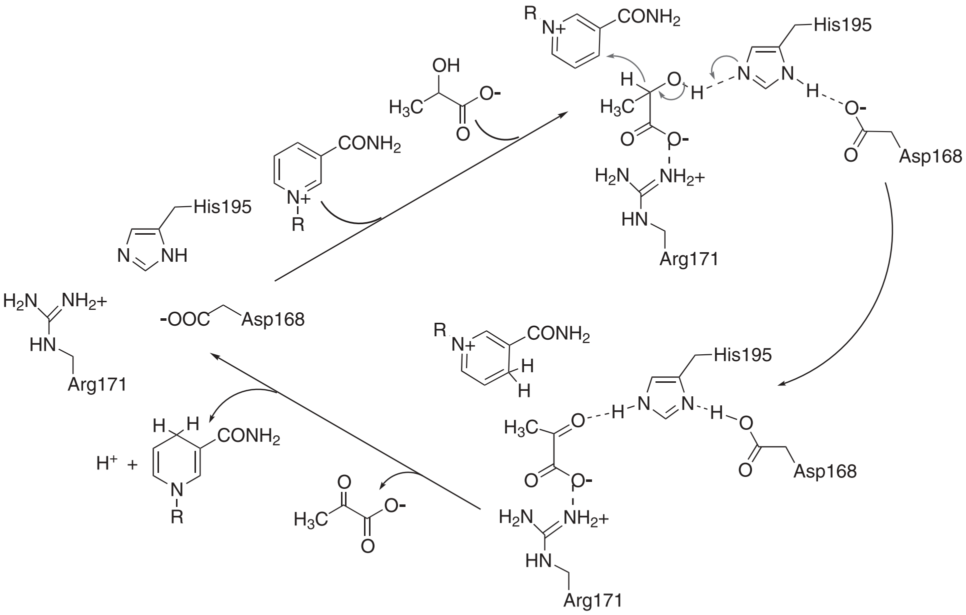


Figure 11.8 Mechanism of redox reaction catalyzed by NAD$^+$ dependent lactate dehydrogenase Lactate dehydrogenase (EC 1.1.1.27) is a tetrameric enzyme which catalyzes the reversible redox reaction between L-lactate and pyruvate *via* ordered kinetic sequence. The hydride ion is transferred to the proR side of the 4 position of NAD$^+$. His195 acts as an acid-base catalyst removing the proton from lactate during oxidation. The active site loop (residues 98–110) carries Arg109 which helps stabilize the transition state during hydride transfer and contacts required for the substrate specificity.

1. *Electron transfer*: Many physiological metallic ions such as Heme-Fe(III)/Fe(II), (Fe-S cluster)$^{n-}$/(Fe-S cluster)$^{(n+1)-}$, Mo(VI)/Mo(IV), Hg(II)/Hg(0), which undergo redox reactions, are facultative or obligate one-electron donors-acceptors. Flavosemiquinone is the essential intermediate in these cases where electron transfer occurs· Two semiquinone radicals, the red radical of flavosemiquinone anion (Fl•$^-$) and the blue radical of neutral semiquinone (HFl•), are identified to participate in the electron transferase reactions (Figure 11.9). The stability of the two semiquinone radicals is governed by the amino acid residue of apoenzymes pointing toward N5 of the coenzyme. The —COO$^-$ group of Asp, Glu stabilizes HFl• (e.g. xanthine oxidae with Mo/Fe-S cluster), while —NH_3^+ of Lys stabilizes Fl•$^-$ (lactate oxidase with Zn^{2+}). Imidazole group His, depending on its charge state, may stabilize either blue or red radicals (glucose oxidase).

2. *Dehydrogenation/transhydrogenation*: Dehydrogenation refers to the redox reaction between substrate and NADH/NADPH via flavin coenzymes catalyzed by flavin-containing dehydrogenases, whereas transhydrogenation occurs between flavin mediated NAD(P)H/NAD(P)$^+$ redox reaction in the absence of the substrates. Thus the electron transfer in the NADH/NADP-linked dehydrogenation is usually accompanied by transhydrogenation. The mode of electron transfer for NADH/NADP oxidation or for dithiol oxidation to disulfide, for example, takes place at the C4a—N5 region of the isoalloxazine ring as the major port of entry of electrons into oxidized

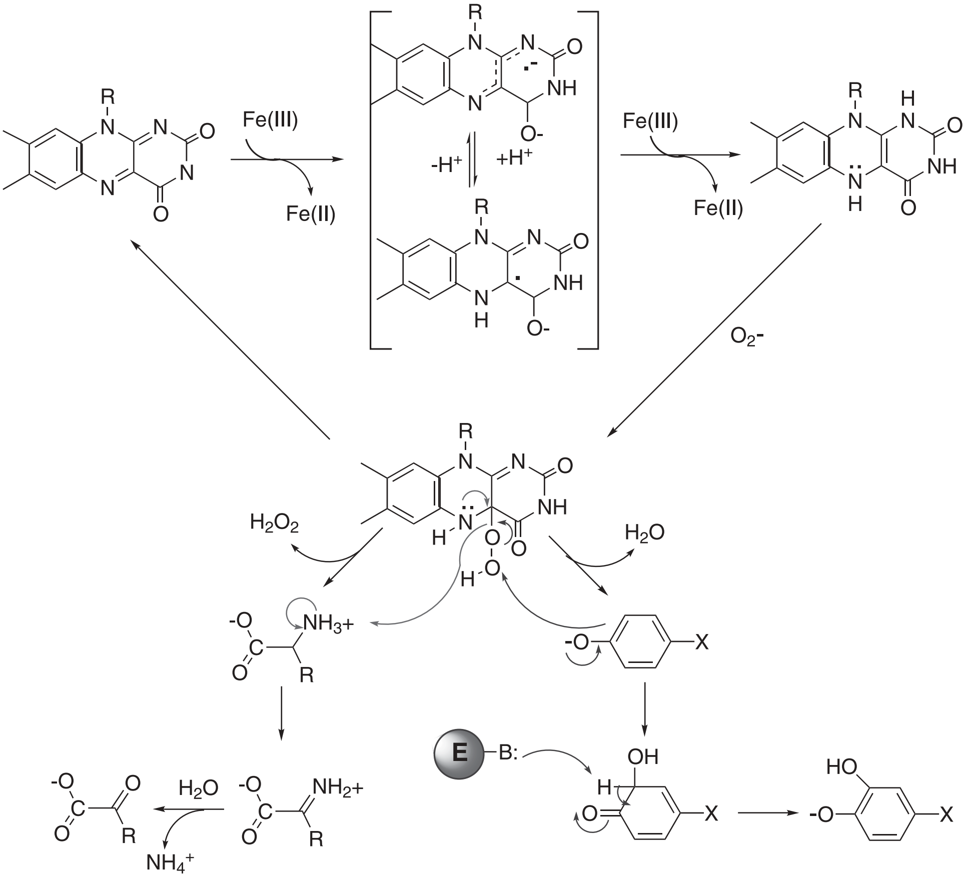


Figure 11.9 Flavoenzyme catalyzed electron transfer and oxidation/oxygenation reactions
The extensive conjugation of the isoalloxazine ring system results in the yellow chromophore (λ_{max} = 450 nm) in the oxidized flavin. Flavin semiquinones are stable radicals, because the unpaired electron is highly delocalized through the conjugated isoalloxazine structure. The neutral semiquinone is blue (λ_{max} = 570 nm) and the flavosemiquinone anion is red (λ_{max} = 480 nm). The reduced 1,5-dihydroflavin isomer, with interrupted conjugation is a leuco (bleached) form. These various redox forms of the flavin coenzymes are amendable to spectrometric monitoring during enzymatic reactions. The electron transfer reaction utilizing cytochrome Fe(III)/Fe(II) (EC 1.9.••.••) is illustrated. The flavin coenzyme activates molecular oxygen to form flavin-4a-hydroperoxide intermediate which reverts back to flavin and H_2O_2 in the amino acid oxidase (EC 1.4.3.2) reaction or oxygenate substrates in the p-hydroxybenzoate monooxygenase (EC 1.14.13.2) reaction.

flavin. The reduced flavin, in the presence of the active site electron acceptor (e.g. lipoamide dehydrogenase, glutathione reductase), passes the reduced equivalent to the active site acceptor. Experimental observations leading to formulation of the reaction mechanism catalyzed by glutathione reductase (EC 1.6.4.2) as proposed/depicted in Figure 11.10 (Pai and Schulz, 1983) are summarized as:

a) Glutathine reductase (GR) catalyzes: GSSG + NADPH + H^+ → 2GSH + $NADP^+$. The enzyme from erythrocyte consists of two identical subunits (subunit weight = 50.6 kDa/FAD). Each subunit contains one FAD, possesses FAD-domain, $NADP^+$/NADPH-domain, central domain and interface domain:

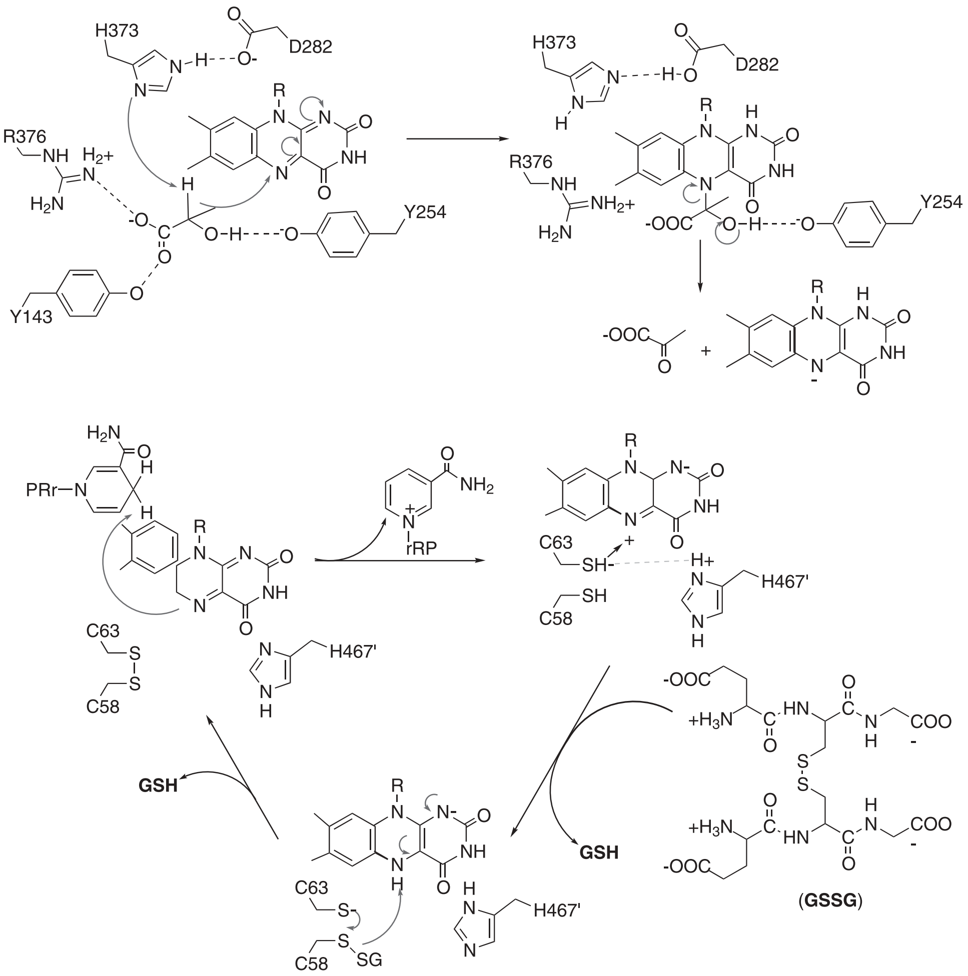


Figure 11.10 Reaction mechanism of glutathione reductase catalysis
The catalytic cycle starts with NADPH binding to the oxidized enzyme with 2′-phosphate group serving as an anchor. Nicotinamide stacks onto the *re*-face of flavin. Two reduction equivalents are transferred from NADPH to the flavin. From the *si*-face of flavin, two electrons flow *via* C4a to Cys63 leading to the opening of the redox-active disulfide bridge. This is followed by the formation of a charge transfer complex between the thiolate anion of Cys63 and reoxidized flavin. The negative charge of Cys63 is assumed to be stabilized by an ion pair interaction with the protonated His467′ to form the half-reduced enzyme. After oxidized glutathione (GSSG) is bound, Cys58 attacks GS-SG to form a mixed disulfide, where the other half of GSSG picks a proton form His467′ and is released. The deprotonated His467′ no longer stabilizes the charge complex. Thus Cys63 proceeds with a nucleophilic attack on the sulfur of Cys58 to restore the redox-active disulfide and releases GSH.

b) Kinetic studies of GR catalysis suggest an operation of the hybrid kinetic mechanism consisting of loops of ping pong and an ordered pathway united by a common NADPH addition step.

c) FAD and NADPH bind to their respective domains. The nicotinamide ring and the isoalloxazine ring are stacked on to each other so that C3 of nicotinamide opposes C4a′ of flavin with three amino acid residues in the close vicinity; Lys66 is salt-bridged to Glu201 and His467′ (i.e. His467 from the neighboring subunit).

d) The critical role of His467 is implicated by chemical modification (ethoxycarbonylation and photooxidation) and site-directed mutagenesis (H467Q).

e) Reductive carboxymethylation identifies the active site disulfide, Cys58-Cys63 as the reduce-equivalent acceptor and substrate site. The redox-active disulfide is strategically located between the isoalloxazine ring of FAD and disulfide linkage of oxidized glutathione (GSSG). The spectroscopic study implicates the formation of a charge transfer complex between its thiolate anion and flavin.

f) Three reaction intermediates of the catalytic cycle have been trapped in the crystal by X-ray crystallographic analysis. They are the half-reduced enzyme (EH_2), EH_2—$NADP^+$ and mixed disulfide between enzyme and glutathione.

Other substrates for flavo-dehydrogenases include structures with electron-withdrawing (activating) groups next to the position of dehydrogenation. Such reactions appear to be initiated by abstraction of the relatively acidic α-hydrogen atom by a basic residue (e.g. His), thus involving a carbanion of the substrate as an intermediate or a transition state.

3. *Oxidation/oxygenation*: When molecular oxygen is the reducible substrate for dihydroflavin reoxidation, reactions involving net reduction of O_2 by one-electron, two-electron and four-electron are known, possibly all arising from an initial common reaction pathway leading to formation of the key intermediate, flavin-4a-hydroperoxide. In flavoenzymes, the fate of flavin-4a-hydroperoxide is distinct. In oxidases, the intermediate breaks down intramolecularly to H_2O_2 and oxidized flavin (e.g. amino acid oxidase). In oxygenases (e.g. p-hydoxybenzoate monooxygenase), the hydroperoxide undergoes O—O bond fission and substrate oxygenation (Figure 11.8). In phenolic hydroxylation, the flavin-4a-hydroperoxide intermediate acts as electrophilic oxygen by donating the distal oxygen to a phenolic carbanion equivalent to generate the oxygenated product. FAD-linked S/N-oxygenases act to deliver electrophilic oxygen to the bound S/N atom of the sulfide/amine substrates. The insertion of an oxygen into a carbon–carbon bond as a ketone is converted to a lactone, in which the distal peroxide oxygen of the flavin-4a-hydroperoxide acts as a nucleophilic equivalent.

11.4.6.2 Hydrolase: chymotrypsin and arginase. Chymotrypsin (EC 3.4.21.1) is an endopeptidase catalyzing hydrolytic cleavage of amide and esteric linkages, in which the carbonyl donor is an aromatic amino acid (Blow, 1976; Perona and Craik, 1997). The enzyme is mechanistically representative of a family of serine proteases having reactive serine residues with a pH optimum around neutrality. The experimental observations leading to the formulation of the proposed/depicted mechanism for chymotrypsin catalysis (Figure 11.11) can be summarized briefly as:

a) Steady state and transient kinetic studies are consistent with the kinetic sequence:

$$E + A \underset{k_2}{\overset{k_1}{\rightleftharpoons}} EA \underset{k_4}{\overset{k_3 \nearrow P}{\rightleftharpoons}} EQ \underset{k_6}{\overset{k_5}{\rightleftharpoons}} E + Q$$

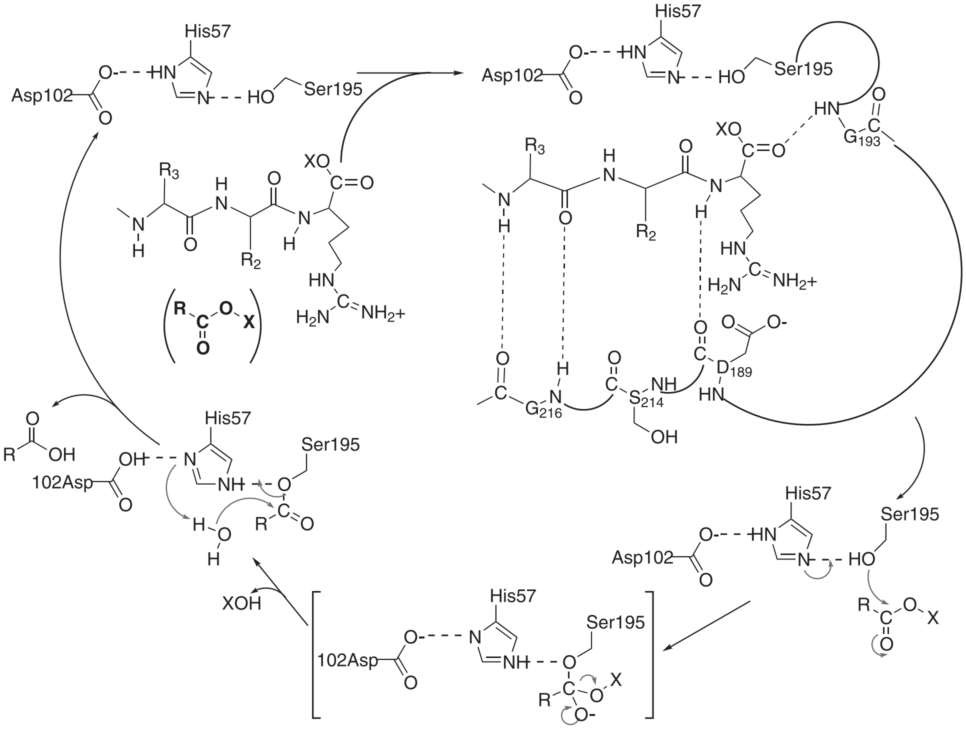


Figure 11.11 Mechanism of chymotrypsin catalyzed hydrolysis
The polypeptide chain of α-chymotrypsin is fold into two domains, each of ~120 amino acids. The two domains are both of the antiparallel β-barrel type, each containing six β-strands. The active site is situated in a crevice between the two domains. Domain 1 contributes two residues in the catalytic triad, His 57 and Asp102 whereas the catalytic Ser195 is the part of domain 2. The enzyme hydrolyzes peptide bond following aromatic side chain at P_1. The P_1 side chain (shown as Arg) of the substrate fits into a crevice on the enzyme surface composed of residues 214–216 on one side and residues 190–192 on the other. These side chains are not shown in the subsequent steps for clarity. The catalysis proceeds in two steps: (1) nucleophilic attack by Ser195 with a departure of the leaving group and the formation of the acyl enzyme *via* tetrahedral transition state and (2) deacylation *via* indirect base catalysis by His57 with a release of the acyl group. Tight binding and stabilization of the tetrahedral transition state is accomplished by hydrogen bond formation between the carboxyl oxygen of the substrate and the main chain NH groups of residues 193 and 195 (not shown) in an oxyanion hole of the enzyme.

notably, $k_{cat} \approx k_3$ $(k_3 \ll k_5)$ for NAc-L-Trp amides, whereas $k_{cat} \gg k_2$ $(k3 \gg k_5)$ for NAc-L-Trp esters. Thus, amides assay acylation step while esters measure deacylation step.

b) pH-Rate profiles are associated with two ionization groups ($pK_1 = 7.16$ and $pK_2 = 8.9$) for amides and one ionization group ($pK = 6.86$) for esters, corresponding most likely to His and Ser for acylation and to His for deacylation steps respectively.

c) Studies on structure-activity relationships reveal that

$$\log(k_{cat,X}/k_{cat,H}) = -2.0\ \sigma \text{ for amides and}$$

$$\log k_{cat} = 2.201\ \sigma^* + 1.012\ E_s + 0.374\ \pi - 2.067 \text{ for esters.}$$

These results suggest that an electrophilic assistance for the acylation and a nucleophilic catalysis for the deacylation.

d) Common to serine proteases, the binding site consists of subsites, S_1, S_2, S_3, S_1', S_2', S_3' corresponding to bound substrates, P_1, P_2, P_3 on the acyl group side/N-terminal side and P_1', P_2', P_3' on the amide or ester side/C-terminal side. A stretch of extended polypeptide backbone chain, Ser214-Trp215-Gly216 forms a typical antiparallel β-pair hydrogen bonded structure with the peptide chain of the substrate.

e) Chemical modifications implicate the importance of Ser195 and His57 to the enzymatic activity.

f) The crystal structure of α-chymotrypsin shows that the active site is a crevice located between the two domains in which one domain contributes two residues His57 and Asp102 and the other domain, Ser195 of the catalytic triad.

g) X-ray crystallographic analysis indicates that the catalytic Ser195 is hydrogen bonded to the imidazole of His57, which in turn is hydrogen bonded to the carboxylate of Asp102, forming a charge relay system.

h) The X-ray crystal structure of the acyl chymotrypsin intermediate has been elucidated.

i) The low barrier hydrogen bond formation (Cleland *et al.*, 1998) between His57 and Asp102 increases the basicity of His57, which in turn enhances its reactivity as a base in removing the proton from Ser195 and lowers the energy of the transition state for forming the tetrahedral intermediate.

j) Thermodynamic calculations indicate that serine proteases work by providing electrostatic complementarity to the changes in charge distribution occurring during the catalyses. The anionic transition state is stabilized by the positively charged His, which in turn is stabilized by Asp. One hydrogen bond from the oxyanion hole (residues 193–195) also contributes ~21 kJ mol^{-1} to the stabilization of the tetrahedral transition state.

Some hydrolases require metal ions for catalyses. Manganese and zinc ions are the most common cofactors in metalloenzymes that catalyze hydrolysis and hydration reactions. The Mn^{2+} is characterized as hard and tends to prefer hard ligands such as carboxylate oxygen's of Asp/Glu, carboxamide oxygen's of Asn/Gln and sometimes imidazole nitrogen of His, to form usually square pyramidal or trigonal bipyramidal with a coordination number of 5 (sometime octahedral with a coordination number of 6). Notably the sulfur of Cys, considered a soft ligand, has not been observed to coordinate Mn^{2+}-metalloenzymes. The Zn^{2+} is characterized as a metal ion of borderline hardness and tends to complex with more diverse arrays of soft and hard ligands to form usually tetrahedral or distorted tetrahedral geometry with a coordination number of 4. The role of these metal ions is to stabilize and position a reactive hydroxide ion, thereby ensuring that an activated nucleophile is available for catalysis. Arginase (EC 3.5.3.1) is a manganese-containing metallogenzyme that catalyzes the hydrolysis of the arginine side chain to form ornithine and urea (Christianson and Cox, 1999). Enzymological data in conjunction with the crystal structure suggests a metal-activated hydroxide mechanism for arginine hydrolysis (Figure 11.12). The binding study of tetrahedral arginine analog inhibitor, (S)-2-amino-6-boronohexanoic acid hydrate, $[H_2O \cdot (HO)_2B(CH_2)_4CH(NH_2)COO^-]$ is consistent with the formation of a tetrahedral intermediate during the hydrolytic reaction.

11.4.6.3 *Transferase, isomerase and lyase: pyridoxal enzymes.*

Pyridoxal-5′-phosphate (PLP) is an obligate coenzyme for the great majority of enzymes catalyzing

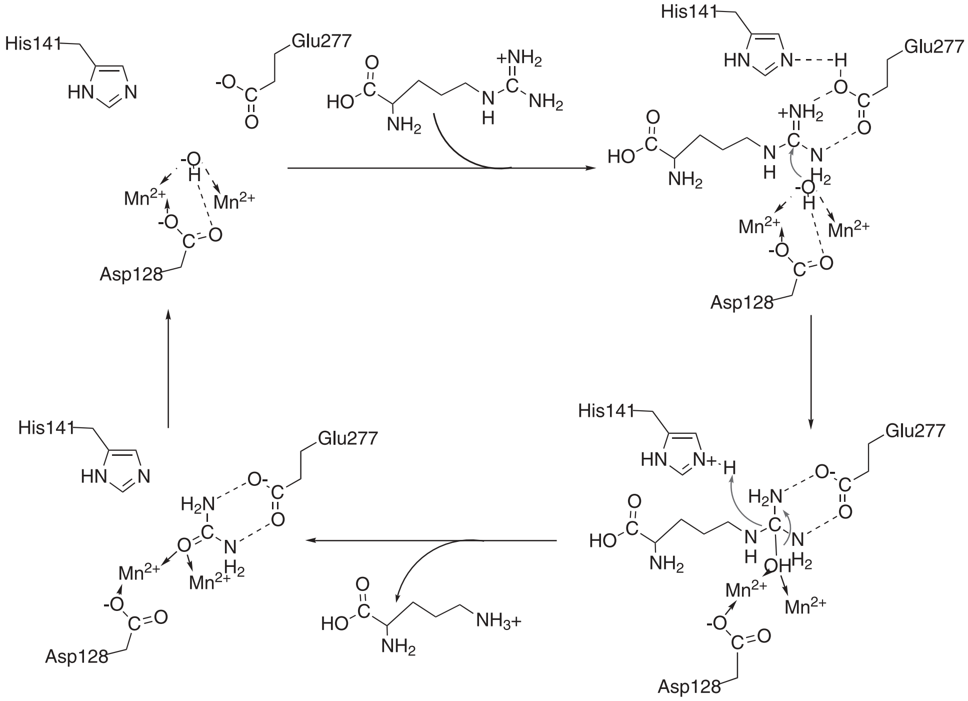


Figure 11.12 Mechanism of arginase catalyzed hydrolysis
Arginase is a homotrimeric metalloenzyme. A binuclear manganese cluster (Mn^{2+}—Mn^{2+} cluster with internuclear separation of 3.3 Å) is required for full catalytic activity. The metal ion situated deep in the base of the active site cleft is designated Mn_A^{2+} and is coordinated by His101, Asp-124, Asp-128, Asp-232 and a bridging hydroxide ion (the hydroxide ion also hydrogen bonded to Asp128) with square pyramidal geometry. Metal ion, Mn_B^{2+} is coordinated by His126, Asp124, Asp232, Asp234 and the bridging hydroxide ion in a distorted octahedral geometry. Both metal ions coordinate to polarize the hydroxide effectively. The side chain of Glu277 which is located at the base of the active-site cleft adjacent to Mn_A^{2+} for salt-bridges with the substrate guanidinium group and placing the scissile guanidinium carbon directly over the metal-bridging hydroxide ion for the nucleophilic cleavage. After proton transfer to the leaving amino group of ornithine, the tetrahedral intermediate collapses releasing ornithine and then urea. Located on one wall of the active-site cleft, His141 may facilitate product release by serving as a proton shuttle.

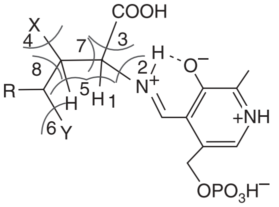
various chemical transformations (Eliot and Kirsch, 2004) at the α-, β- or γ-carbons of α-amino acids (Table 11.13), such as alanine-oxoacid transferase (EC 2.6.1.12), amino acid racemase (EC 5.1.1.10) and histidine carboxy-lyase (EC 4.1.1.22).

All pyridoxal enzymes catalyzing chemical transformations of α-amino acids function via:

1. initial imine (pyridoxylidene imino acid) formation;
2. chemical changes involving carbanionic intermediates; and
3. hydrolysis of a product imine.

The reaction specificity results from the nature of the individual enzyme proteins. The role of PLP is to condense with amino acid to form a Schiff base (pyridoxylidene imino acids), which acts as an electron sink to stabilize carbanionic intermediates

TABLE 11.13 Pyridox phosphate mediated chemical transformations of α-amino acids. After condensation of α-amino acids with pyridoxal-5′-phosphate (PLP) to form pyridoxylidene imino acids (structure shown), chemical transformations may occur in several directions depending on the subsequent bond cleavage step(s)



Site indicated	Bond cleavage	Reaction
1	C_α—H	Racemization
1 + 2	C_α—H, C_α—N	Transamination
3	C_α—COOH	Decarboxylation
1 + 4	C_α—H, C_β—X	Elimination or replacement of α-H and β-X
5 + 6	C_β—H, C_γ—Y	Elimination or replacement of β-H and γ-Y
7	C_α—C_β	α,β-cleavage (and condensation) of carbon chain
8	C_β—C_γ	β,γ-cleavage of carbon chain

developed during enzymatic catalysis. In pyridoxal enzymes, the PLP coenzyme is bound to the ε-NH_2 of the active site Lys (Figure 11.7). The reactions at C_α of α-amino acids can proceed by racemization, transamination or carboxylation, depending on which of the C_α substituents is labilized (Figure 11.13).

Model studies (pyridoxal catalyzed conversion of α-amino acid to oxo-acid) indicates that the prototropic shift is in the aldimine ←→ ketimine tautomerization, and this step can be greatly accelerated by general acid–base catalysis. Aspartate (:2-oxoglutarate) aminotransferase (EC 2.6.1.1), which catalyzes transamination between Asp and 2-oxoglutarate (oxaoacetate and Glu), is the most extensively studied representative PLP enzyme. The enzyme is a homodimer containing one PLP molecule per subunit. Experimental observations pertaining to apartate aminotransferase are:

a) The transamination proceeds via the ping pong bi bi kinetic mechanism.

b) The PLP coenzyme forms imine linkage to the ε-amino group of the essential Lys258.

c) It is possible to achieve complete conversion of the enzyme into the enzyme–substrate complex, which displays absorption bands (λ_{max}) at 492, 430, 362 and 330 nm with the prominent 492 nm band attributed to the quinoid intermediate (Figure 11.7).

d) The ε-amino group of Lys258, which is released in the enzyme–substrate imine formation, serves as the base to abstract the α-hydrogen from the substrate.

e) The activity of the inactive mutant, K258A can be restored by exogenous amine, suggesting that Lys258 also acts as an acid–base catalyst for the 1,3 prototropic shift in aldimine/ketimine interactions.

f) Both the pyridine N-atom and the imine N-atom are protonated.

g) The proton is delivered from the *si* face of the enzyme bound planar aldimine, yielding proS pridoxamine phosphate.

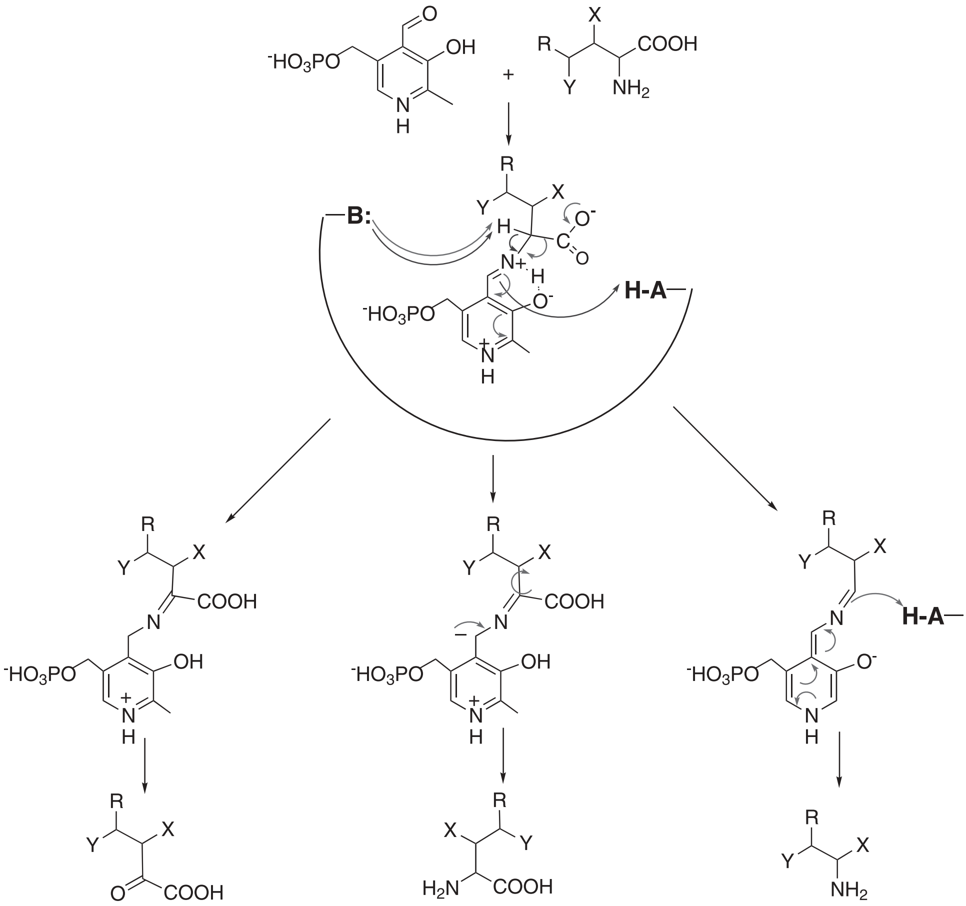


Figure 11.13 Reactions at α-carbon of α-amino acids catalyzed by pyridoxal enzymes
All three substituents at C_α are subject to labilization in the three types of α-carbon reactions. The hydrogen is labilized in recemization reactions, the amino group is labilized in the transamination and the carboxyl group is labilized in decarboxylation. α-Amino acid condenses with pyridoxal phosphate to yield pyridoxylidene imino acid (an aldimine). The common intermediate, aldimine and distinct ketimines leading to the production of oxo-acid (in transamination), amino acid (in racemization) and amine (in decarboxylation) are shown. The catalytic acid (H-A–) and base (–B:) are symbolic; both can be from the same residue such as Lys258 in aspartate aminotransferase.

The best known PLP requiring amino acid racemase is alanine racemase (EC 5.1.1.1), which produces a key building component of bacterial peptidoglycan, D-alanine from L-alanine. The reaction can be envisioned as an abstraction of α-hydrogen, followed by delivery of a proton back to the imine intermediate after C—C bond rotation. The racemization occurs if the enzyme delivers a proton back to either face of the planar imine intermediate. The enzyme appears to catalyze both inversion and retention of configuration of the substrate with a similar probability producing a racemic mixture. The α-decarboxylation catalyzed by amino acid decarboxylation has been shown to proceed in retention with the decarboxylation as the rate-limiting step. The aldimine is hydrolyzed to give amines.

After an initial α-hydrogen abstraction, transformations at β-carbon or γ-carbon of α-amino acids may occur. When α-amino acid substrates possess a substituent at either β- or γ-position, they can function as good leaving groups such as COO^-, OH(CH3) or SH. Either elimination or replacement of the substituent yields corresponding oxo acids. In the case of C_γ reactions, labilization of β-hydrogen to form a β-carbanion intermediate is necessary.

11.4.6.4 Ligase: Aminoacyl-tRNA synthetase. The aminoacylation of tRNA is catalyzed by a family of 20 aminoacyl-tRNA synthetases (EC 6.1.1.••), each specific for one amino acid and one or more isoaccepting tRNA. The aminoacylation (or charging) of tRNA by two classes of aminoacyl-tRNA synthetases (class I and class II aRS's) is carried out in two steps. The amino acid is activated by attacking a molecule of ATP at the α-phosphate, giving rise to aminoacyl-adenylate in the first step, and the amino acyl group is transferred to the 3′-terminal ribose of a cognate tRNA yielding an aminoacyl-tRNA in the second step, as depicted in Figure 11.14.

Following features characterize the aRS catalyses:

1. *Adenylylation of amino acids*: Crystal structural analyses of aRS complexed with substrates, analogs and intermediates (e.g. enzyme-bound activated amino acid) provide a wealth of information concerning aRS catalyzed activation of amino acids. Both classes of aRS's employ the oxygen and nitrogen atoms of the peptide-backbone loop elements specifically to select the adenine base against guanine or pyrimidine bases of riboside triphosphates. The two signature sequences, HIGH and MSK of class I aRS's, interact ATP. Two side chains, Gly in HIGH and Lys in MSK, appear to have an overriding importance. Gly is implicated in the interaction with the nucleotide and Lys is implicated to participate in the transition state stabilization. In class II enzymes, ATP is fixed by the backbone portion of the motif 2 loop. The in-line nucleophilic attack by the carboxylate of amino acid at α-phosphate yields an oxyphosphorane (pentacoordinate) transition state, having bipyramidal geometry with the attacking and departing atoms at apical positions. In the case of TyrRS (class I), mobile loops on either side of the active site provide four positive charges (Lys82, Arg86, Lys 230 and Lys 233) that seize the pyrophosphate (β- and γ-phosphates) and pull it from the α-phosphate as the ribose is sucked in the opposite direction. It is likely that class I aRS's use strain and induced fit mechanisms in amino acid adenylylation (Carter, 1993).
2. *Aminoacylation of tRNA*: The acceptor arm and the 3′-terminal –CCA of tRNA are bound by both classes of aRS's in a mirror-symmetric fashion with respect to each other. Class I aRS uses an α-helix and a loop to interact with base pairs (bps) in the minor groove of the acceptor stem of tRNA. As the result, the 2′-OH of the 3′-terminal ribose is brought into a position to attack the carbonyl of the aminoacyl adenylate. By contrast, class II aRS uses the long loop of motif 2 to interact with the discriminator base and the first bp in the major groove. The CCA end is bent into the active site and brings the 3′-OH of the 3′-terminal ribose in the reactive position. Mechanistically, the two classes of an aRS's differ in that the class I aRS transfers aminoacyl group to the 2′-OH of the terminal ribose of CCA, whereas the class II aRS aminoacylates the 3′-OH of the terminal ribose of CCA of the cognate tRNA.

Specificity of aminoacylation: The faithful translation of genetic information depends on the correct matching of amino acids to their cognate tRNA that is accomplished by aRS catalysis. The degeneracy of the genetic code versus the nearly general provision of only one aRS per amino acid poses an intriguing question about the specificity of aRS cataly-

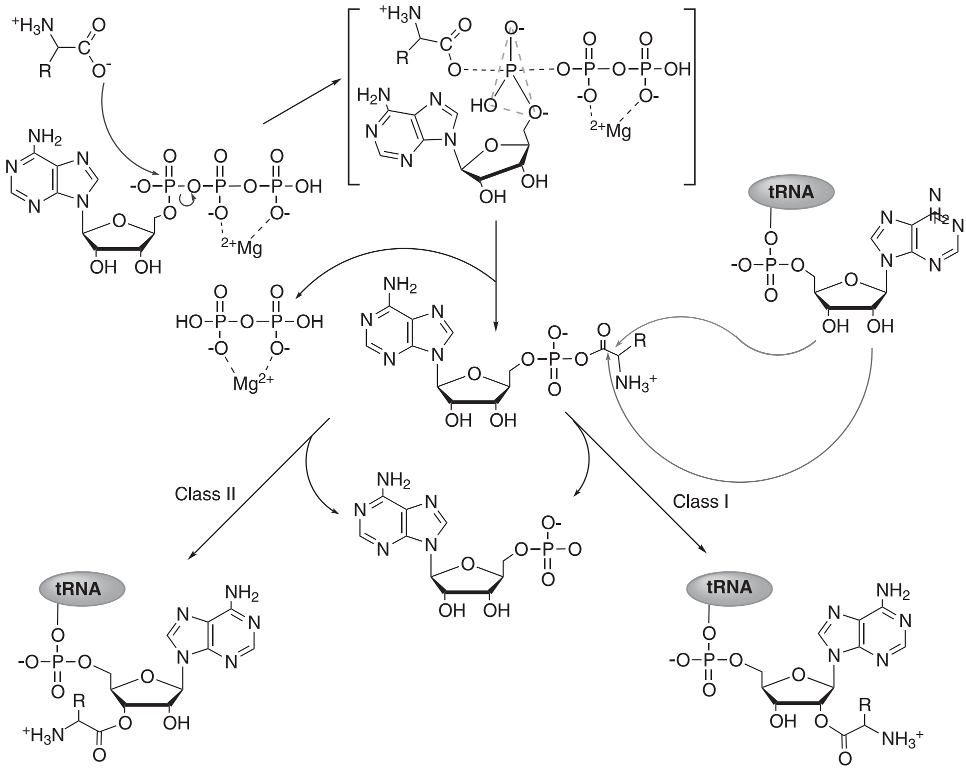


Figure 11.14 Aminoacylation mechanisms catalyzed by aminoacyl-tRNA synthetases
The two classes of aminoacyl-tRNA synthetases (aRS's) differ in the site of aminoacylation. Class I aRS's aminoacylate 2′-OH whereas class II aRS's add amino acids to 3′-OH of the terminal ribose of the 3′-terminal CCA of cognate tRNA. Magnesium ions complexed with ATP to enter the active site of aRS may play a dual role in the activation step by both stabilizing the conformation of the ATP (Mg^{2+} ion bridges the β- and γ-phosphates) and participating in adenylate formation (second Mg^{2+} is found between α- and β-phosphates in some aRS's). In class I aRS, both Lys of MSK and His of HIGH stabilize the bipyramidal oxyphosphorane transition state while R of motif 2 in class II aRS participates in the stabilization of the putative pentacoordinate transition state. The resulting mixed anhydride aminoacyl adenylate is held by the enzyme for the next reaction, i.e. the attack by the 2′-OH (class I) or 3′-OH (class II) of the terminal adenosine at the carbonyl of the aminoacyl adenylate. The amino acid then becomes esterified to the cognate tRNA.

sis. A set of tRNA isoacceptors can be distinguished mainly by the acceptor arm, the anticodon loop and a few bases pairs in the T and D stems. Each group of isoacceptor tRNA possesses a certain identity elements that the cognate aRS can recognize but causes rejection by noncognate aRS. Amino acid and cognate tRNA bindings are probably coupled to protein conformation changes during amino acid activation and aminoacyl transfer processes. To form correct and productive complex, aRS and/or tRNA must undergo unfavorable conformational changes. The complex buries 2500–2700 Å^2 of the contact area, which could provide about 200 kJ/mol, of which about 1/5 is used to stabilize the complex. The difference (~160 kJ/mol) represents a source of free energy for distorting structures of the uncomplexed aRS/tRNA and using these distortions to enforce specificity. In the class I aRS, the MSK loop changes its conformation upon tRNA binding while a flipping loop, which is located between motifs 1 and 2 in the class II aRS, opens up when the correct tRNA is bound. Both classes of aRS's possess dedicated domains that interact specifically with the anticodons of their cognate tRNA and induce extensive conformational changes to maximize the contact area.

11.5 ENZYME REGULATION

11.5.1 Elements of enzyme regulation

The myriad of different enzymatic reactions that occur must be regulated for the proper functioning of a living system. This modulation is achieved by the regulation of key enzymes that control metabolic fluxes (Stadtman, 1966; Hammes, 1982). Enzyme regulation is carried out by one of the following strategies.

1. *Quantity of enzyme*: The enzyme concentration in cells is regulated at the synthetic level by genetic control, which may occur positively or negatively. Alternatively, the control can be exerted by the specific degradative pathways. The genetic control of enzyme (protein) synthesis and specific protein degradation will be considered in the Chapters 12 and 13.
2. *Catalytic activity*: The regulation of enzyme activities is achieved in two modes, namely switching on-and-off and tuning up-and-down. Covalent modifications of enzymes effectively switches their activities on or off, whereas noncovalent effectors tune enzyme activities up or down by affecting their kinetic parameters.
3. *Availability of enzyme/substrate*: The compartmentation/solubility of an enzyme and/or its substrate/cofactor is another form of controlling the activity of the enzyme. Metabolic activities in cells are compartmentalized into different cellular structures and organelles (Table 3.2). The enzymes that catalyze sequential reactions in the pathway may be bound to the structural element of the cell and form a stable multienzyme complex known as metabolon (Srere, 1987). The multienzyme complexes are advantageous for coordination of component enzyme activities, facilitation of metabolite channeling along the metabolic pathway and maximization of the regulatory effect of the key enzyme. In eukaryotics, membranes provide such compartmentalization. For example, the electron transport system is confined to mitochondria, protein synthesis on ribosomes and hydrolytic enzymes to lysosomes. The membrane also controls the enzyme/metabolic activities by modulating the flux of essential metabolites known as metabolite shuttles. For example, fatty acid oxidation is confined within mitochondria, whereas fatty acid synthesis takes place in cytosol. Carnitine functions as an acyl carrier in the fatty acid shuttle across the membrane barrier. Metabolic enzymes concerned with an individual pathway are often localized in one particular cellular compartment.

11.5.2 Covalent modifications of enzymes and cascade effect

The covalent modification is an important control mechanism of enzyme activities. Some enzymes are synthesized in inactive precursor forms known as proenzymes or zymogens, which are activated at a physiologically appropriate time and place. For example, the activation of chymotrypsinogen occurs in stages. While chymotrypsinogen may harbor the active site Asp···His···Ser charge relay system of the active enzyme, the zymogen is inactive (the activity is ~10^6 fold less than chymotrypsin) because the substrate binding pocket is not properly formed. An initial cleavage of the R15–I16 bond catalyzed by trypsin creates a new positive charge at the α-amino group of I16. An electrostatic interaction between this positive charge and that of the side chain of D194 helps orientating other parts of the molecule, such as the side chains of R145 and M192 to form the substrate-binding pocket. This yields a fully active, two-chain enzyme known as π-chymotrypsin. Subsequent autocatalytic cleavage at L13–S14 bond yields δ-chymotrypsin, and finally α-chymotrypsin by cleaving Y146–T147 and N148–A149 bonds (Gertler *et al.*, 1974).

However, in most cases, covalent modifications can be either activation or deactivation and are normally energy-dependent and reversible, but catalyzed by separate enzymes. These include phosphorylation/dephosphorylation, adenylylation (nucleotidylation)/deadenylylation and ADP-ribosylation (Stadtman and Chock, 1978). This cyclic interconversion of key enzymes between covalently modified and unmodified forms is a mechanism of singular importance in cellular regulation. The interconversion of an enzyme between the active and inactive forms involving separate modification and demodification (i.e. different converter enzymes) is a dynamic process that may lead to a steady state in a cascade system. The covalent interconversion of regulatory enzymes is characterized by:

- the formation of stable forms of the modified/demodified enzymes;
- different converter enzymes are involved in the cyclic cascade;
- it is energy-dependent, but less energy is required than the *de novo* enzyme synthesis via genetic control;
- it can maintain partial activities in multienzyme complexes;
- it can respond to a greater number of stimuli by an increase in the number and/or type of effector sites;
- it exhibits greater flexibility in the control patterns; and
- it can achieve greater amplification of regulatory signals.

Phosphorylation/dephosphorylation is the principal regulatory mechanism for enzyme activities (Krebs and Beavo, 1979) as well as for varied cellular activities such as receptor interaction, hormonal function, cell cycle and mobility, oncogenesis, and signal transduction. Some target enzymes affected by phosphorylation/dephosphorylation (↑ for increase and ↓ for decrease in activities) are: glycogen synthase (↓/↑), glycogen phosphorylase (↑/↓), fructose –1,6-*bis*phosphatase (↓/↑), L-type pyruvate kinase (↓/↑), PDC (↓/↑), hormone-sensitive lipase (↑/↓), cholesterol esterase (↑/↓), acetyl CoA carboxylase (↓/↑), tyrosine hydroxylase (↑/↓) and RNA polymerase (↑/↓). The major phosphorylation sites are Sere, Thr and Tyr with Cys, His, Lys, Gln, and Arg being possible sites. The cyclic phosphorylation/dephosphorylation cascade consists of protein kinases (PrK), phosphoprotein phosphatase (PPrP) and associated cofactors as well as effectors. Figures 11.15 and 11.16 illustrate regulation of glycogen metabolism (glycogen synthase and glycogen phosphorylase) by monocyclic (Figure 11.15) and multicyclic (Figure 11.16) phosphorylation/dephosphorylation cascade respectively.

A theoretical analysis of cyclic cascades (Chock *et al.*, 1980) reveals that:

- they possess a great capacity for signal amplification, i.e., a relatively small fractional activation of the converter enzyme can promote a large change in the interconvertible enzyme;
- they can regulate the amplitude of the maximal response that an interconvertible enzyme can achieve with saturation concentration of an effector;

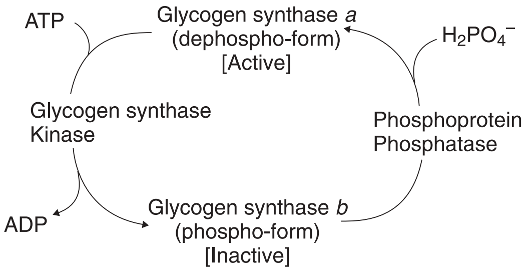

ATP
Glycogen synthase *a*
(dephospho-form)
[Active]
$H_2PO_4^-$
Glycogen synthase
Kinase
Phosphoprotein
Phosphatase
ADP
Glycogen synthase *b*
(phospho-form)
[Inactive]

Figure 11.15 Regulation of glycogen synthase by phosphorylation/dephosphorylation
The active form of glycogen synthase a is the dephosphorylated form which is inactivated by the phosphorylation of two Ser. Glycogen synthase is regulated by the monocyclic phosphorylation/dephosphorylation cascade in a manner reciprocal to that of glycogen phosphorylase.

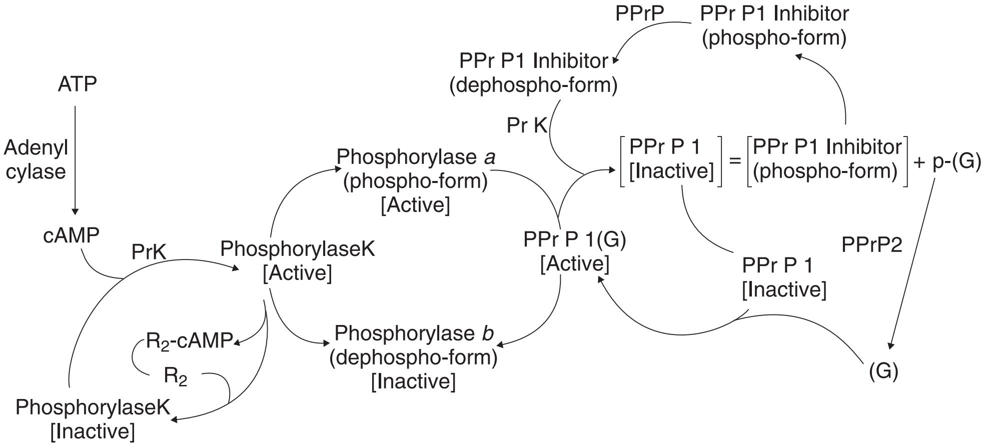

PPrP
PPr P1 Inhibitor
(phospho-form)
PPr P1 Inhibitor
(dephospho-form)
ATP
Adenyl
cylase
Pr K
PPr P 1
[Inactive]
=
PPr P1 Inhibitor
(phospho-form)
+ p-(G)
Phosphorylase *a*
(phospho-form)
[Active]
cAMP
PrK
PhosphorylaseK
[Active]
PPr P 1(G)
[Active]
PPr P 1
[Inactive]
PPrP2
R_2-cAMP
R_2
Phosphorylase *b*
(dephospho-form)
[Inactive]
(G)
PhosphorylaseK
[Inactive]

Figure 11.16 Regulation of glycogen phosphorylase by phosphorylation/dephosphorylation
The major enzymatic system in the regulation of glycogen phosphorylase (i.e. phosphorylase) by the multicyclic phosphorylation/dephosphorylation cascade is shown. Abbreviations used are: Pr, protein; PrK, Protein kinase; phosphorylaseK, phosphorylase kinase (C_2R_2 where C_2 and R_2 are dimeric catalytic and regulatory subunits respectively); PPrP, phosphoprotein phosphatase; (G), G-subunit of phosphoprotein phosphatase and p-(G), phopsho-G-subunit.
Adrenalin activates adenylate cyclase which synthesizes adenosine-3′,5′-cyclic monophosphate (cAMP), an activator of PrK. The enzyme (cAMP-dependent PrK) phosphorylates Ser and/or Thr (with consensus sequence of Arg-Arg-X-Ser/Thr-Y) of phosphorylase kinase consisting of C_2R_2. The binding of cAMP causes the dissociation of active catalytic monomers which utilizes ATP to phosphorylate phosphorylase b to the active phospho-phosphorylase a. The phosphorylation occurs at Ser14 of phosphorylase and requires Ca^{2+}. The dephosphorylation of the active phospho form to the inactive dephospho form is catalyzed by PPrP1 which becomes active when complexed with G-subunit. The complexation of PPrP1(G) with its inhibitor releases phospho-(G) which is dephosphorylated to G-subunit by the action of PPrP2.

- they can modulate the sensitivity of interconvertible enzyme modification to changes in the concentration of allosteric effectors;
- they can adjust the specific activity of the interconvertible enzyme according to fluctuation in the intracellular concentrations of various metabolites;
- they show extreme flexibility with respect to allosteric regulations; and

- they can serve as rate amplifiers capable of responding rapidly in milliseconds to changes in metabolite levels.

11.5.3 Control of enzyme catalytic activity by effectors

The enzymatic activity is increased/decreased in the presence of an activator/inhibitor, which is the simplest form of controlling catalytic activity of enzymes. Enzyme activation/inhibition is the kinetic effect. Their combined effects may give rise in various forms of metabolic regulations and provoke different explanations, some of which are considered here.

11.5.3.1 Biochemical oscillation. The availability or dynamic flux of substrates/cofactors may also cause fluctuation in the operational enzyme activities. For example, the coordinated regulation by the coupled effect of metabolites acting as activators and inhibitors gives rise to the periodic response (cyclic fluctuation) of the product or measurable intermediates known as oscillatory effect (Chance *et al.*, 1973). Biochemical oscillation or biorhythmicity is also observed in the signal transduction systems (Myer and Stryer, 1988; Berridge, 1990). The necessary conditions for oscillations can be stated as:

- For an oscillating metabolic pathway, there are two reactions or two segments of reactions (and therefore the enzymes that catalyze these reactions), which are subject to metabolite activations/inhibitions. The enzymes involved are called oscillophors.
- These two reactions (segments of reactions), which are self-regulating as well as cross-coupled regulating.
- The two self-regulatory terms must be of opposite characters and the two cross-coupling terms must also be of opposite characters.
- The sum of the self-regulating terms must be net positive.
- The magnitude of the product of the cross-coupling terms must be greater than the magnitude of the product of the self-regulating terms.

For example, the two segments of reactions (catalyzed by phosphofructokinase, PFK and glyceraldehydr-3-phosphate dehydrogenase, Gly3PDH/phosphoglycerokinase, PGK) and their effectors responsible for the oscillating utilization of glucose in yeast cultures via the glycolytic pathway are:

Oscillophor	Inhibitor	Activator
PFK	ATP	ADP, fructose -1,6-*bis*phosphate
Gly3P DH/PGK	3PGA	ADP

11.5.3.2 Pasteur effect versus Crabtree effect. Classical examples of the Pasteur effect (inhibition of glycolysis by respiration) and Crabtree effect (inhibition of respiration by glycolysis) can be traced to the regulations mediated via changes in the concentrations of metabolites. The inhibition of glycolysis by oxygen probably mediates through the changes in the concentration of metabolites, which also regulate the activity of phosphofructokinase, PFK (Tejwani, 1978). Facilitation of the PFK reaction during anoxia has been demonstrated in almost all of the tissues in which the Pasteur effect is observed. The activation of PFK is also associated with the activation of hexokinase and pyruvate kinase. The presence of glucose causes a decrease in activities of several respiratory enzymes. For example, in the aerobic respirofermentative metabolism of yeast, pyruvate is the common intermediate channeled through the respirative and fermentative pathways via pyruvate decarboxylase and PDC. Thus the concentration of pyruvate and the relative activities of these two enzymes (pyruvate decarboxylase and PDC) are probably determinants of the relative preference for the respirative versus fermentative pathways. At high concentrations of glucose, the fermentative pathway competes favorably against the respirative pathway (Käppeli and Sonneleitner, 1986).

11.5.3.3 Feedback control. The binding of metabolites, i.e. substrates/products or effectors to enzymes is an important mode of regulation. Feedback inhibition (negative feedback control) is an important example, in which the first committed step in a biosynthetic pathway is inhibited by the ultimate end product of the pathway (Stadtman, 1966). Table 11.14 summarizes different modes of negative feedback controls that have been evolved to accommodate the regulation of divergent metabolic pathways.

Some effectors may act as antagonists to feedback inhibition such as reversal of the inhibitory effect of CTP on apartate transcarbamylase by ATP. In some cases, enzymes are activated by metabolites in the precursor substrate activation when an enzyme catalyzing a key metabolic step is activated by a precursor metabolite. For example, *Saknibekka typhimurium* phosphoenolpyruvate carboxylase is activated by fructose-1,6-*bis*phosphate.

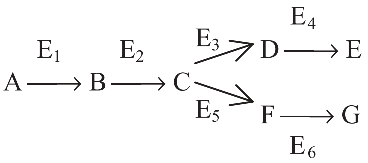
11.5.3.4 Allosterism and Cooperativity. The inhibition/activation by substrates/products is perhaps the most common form of controlling enzyme activities and the measurable kinetic effect is generally linear (double reciprocally transformed, otherwise hyperbolic), though nonlinear (double reciprocally transformed, otherwise nonhyperbolic) situations exist. Similarly, some regulatory enzymes give rise to linear (hyperbolic) kinetics, though many of them exhibit nonlinear (nonhyperbolic) effects. Such behavior of enzymes to regulation is described by allosterism and cooperativity via multiple binding sites on enzymes (Monad *et al.*, 1965; Koshland *et al.*, 1966; Ricard and Cornish-Bowden, 1987) in which effectors bind to the sites (allosteric sites) distinct from the catalytic sites. The allosterism is a mechanistic term referring to the binding of effectors to different sites, while the cooperativity is an operational term concerning interaction between binding sites with respect to their ability to sequential binding of effector molecules. The two molecular models and their treatments have been described in the previous chapter. If the binding of the effector to the first site facilitates binding to the subsequent sites, the interaction is positive cooperativity. On the other hand, if the binding of the first site inhibits the binding to the subsequent sites, the interaction is negative cooperativity. The interactions between binding sites for identical effectors are termed homotropic interactions, which have been discussed in section 10.3. Those between binding sites for dissimilar effectors are called heterotropic interactions, which are typified by the effects of inhibitors and activators.

According to the symmetry model (Monad *et al.*, 1965), an oligomeric allosteric enzyme binds activator A and substrate S to its relax state (R state) and inhibitor I to only the tight state (T state). The fractional saturation for substrate Y_s, i.e. the fraction of substrate binding sites occupied is expressed by

$$Y_s = \frac{\sum_{i=1}^{n}\sum_{j=0}^{n} i[R_{i,j}]}{\left(\sum_{i=1}^{n}\sum_{j=0}^{n} i[R_{i,j}] + \sum_{k=0}^{n}[T_k]\right)}$$

TABLE 11.14 Modes of feedback inhibitions

A branched pathway:



in which A is converted to E and G in a sequence of steps catalyzed by enzymes E$_{1\text{-}6}$, can be regulated by a number of different mechanisms. The first common step is taken as A Æ B catalyzed by E$_1$, which is subject to feedback inhibition

Feedback control	Mechanism	Example
Enzyme multiplicity	The first common step is catalyzed by two or more isozymes, one is inhibited by E and the other by G	*E. coli* 2-Keto-3-deoxy-D-arabinoheptonate synthetase: one isozyme is inhibited by Phe and the other by Tyr
Concerted	The first common step is not susceptible to feedback inhibition by excess of only one end product, E or G but is inhibited by simultaneous presence of excess of E and G.	*Rhodopseud. capsulatus* Aspartokinase is inhibited by excess amount of Lys and Thr but not by Lys or Thr alone.
Cooperative	An excess of any of end products causes a partial inhibition of the first common step, whereas the simultaneous excess of two or more end products results in a greater inhibition	Glutamine phosphoribosylpyrophosphate amidotransferase is inhibited by either of GMP/IMP or AMP/ADP but to a greater extent by their combined presence.
Cumulative	The first common step enzyme is inhibited indepently (residual activity, A_i) by each of the end products at saturation conc. The simultaneous presence of the saturation conc. of two or more end products results in the ultimate inhibition such that the total residual activity, $A_t = \Pi A_i$.	Glutamine synthetase is subject to cumulative feedback inhibition by the end products, carbamyl phosphate, Trp, AMP and CTP.
Heterogeneous metabolic pool	A single inhibitory site for end products exists, but the binding is weak that the site is never fully occupied. The effective inhibitor conc. is the sum of the conc. of all of the end products.	*B. subtilis* purine mononucleotide pyrophosphorylase: inhibition by guanine, hypoxanthine and xanthine.

in which the subscripts i, j, and k denote the respective numbers of S, A and I molecules that are bound to an oligomer, i.e. $R_{i,j} = RS_iA_j$ and $T_k = TI_k$. By analogous treatment (section 10.3), i.e. $\alpha = [S]K_R$ and $C = 0$ (S only binds to R), it gives:

$$Y_s = \frac{\left(\sum_{j=0}^{n}[R_{0,j}]\right)\alpha n(1+\alpha)^{n-1}}{n\left\{\left(\sum_{j=0}^{n}[R_{0,j}]\right)(1+\alpha)^{n}+\sum_{k=0}^{n}[T_k]\right\}} = \alpha(1+\alpha)^{n-1}\Big/\left\{L'+(1+\alpha)^{n}\right\}$$

The expression is identical to that for the homotropic interactions, except that L is replaced by an apparent allosteric constant L′, defined as:

L′ = (sum of different complexes of T states with I)/(sum of different complexes of R states with A):

$$= \frac{\sum_{k=0}^{n}[T_k]}{\sum_{j=0}^{n}[R_{0,j}]}$$

If we define $\beta = [I]/K_I$ and $\gamma = [A]/K_A$, where K_I and K_A are the dissociation constants for bindings of I and A to the T and R states (note: dissociation constants are used to define β and γ according to the practice in enzyme kinetics, whereas α is defined as $\alpha = [S] \cdot K_R$, in which K_R is the association constant) respectively, it can be shown that

$$L' = L\{(1+\beta)^n/(1+\gamma)^n\}$$

and

$$Y_s = \frac{\alpha(1+\alpha)^{n-1}}{L\left\{(1+\beta)^n \big/ (1+\gamma)^n + (1+\alpha)^n\right\}}$$

Therefore, the inhibitor/activator may affect allosteric enzymes by:

1. The inhibitor displays negative heterotropic effect and elevated cooperativity of the substrate: The presence of inhibitor ($\beta > 0$), which only binds to the T state, reduces the binding affinity for the substrate by decreasing the concentration of the R state to which the substrate binds (negative hetereotropic effect). Since the substrate must counter this deficit in the equilibrium concentration of the R state, the cooperativity of the substrate increases in the presence of inhibitor.
2. The activator displays a positive heterotropic effect and suppressed cooperativity of the substrate: The presence of activator ($\gamma > 0$) increases the concentration of the substrate-binding R state and therefore increases the substrate binding affinity for the enzyme (positive hetereotropic effect). The advantage of the equilibrium concentration of the R state decreases the substrate cooperativity in the presence of the activator.

The response of allosteric enzymes to changes in effector concentrations has been assumed to be rapid. However, effectors may induce changes in enzyme activity that occur much more slowly than the rate of the overall catalytic reaction, resulting in a time lag in the response. Such a slowly responding enzyme is termed hysteric enzyme (Goldhammer and Paradies, 1979). Several possibilities account for the hysterisis of enzymes:

a) effector induced conformational changes;

b) displacement of a tightly bound effector by a different effector; and

c) enzyme polymerization-depolymerization.

Alternate models to the differential bindings of the multiple site enzymes are flip-flop mechanisms (Lazdunski *et al.*, 1971) and half-site reactivity (Seydoux *et al.*, 1974) of enzymes.

11.5.4 Structure basis of allosteric regulation: Glycogen phosphorylase

The basic properties of allosterism can be measured by kinetic and ligand binding studies, but the structural basis that modulate these properties can only determined by X-ray crystallographic studies (Perutz, 1990) in conjunction with exploration of allosteric networks and conformational ensembles (Swain and Gierasch, 2006). For example, muscle glycogen phosphorylase (GP) is modulated both by reversible phosphorylation and allosteric interactions (Fletterick and Sprang, 1982) by a number of effectors (Table 11.15). Phosphorylation of Ser14 converts the dephospho-phosphorylase b (GPb) to the phospho-phosphorylase a (GPa) and shifts the T/R equilibrium toward the R state. Glucose returns the equilibrium back to the T state and AMP opposes glucose by shifting the equilibrium to the R state. Inosine monophosphate (IMP) is a weak activator that binds to GPb in the T state (Figure 11.17).

The crystal structures of T-states GPb, GPb-IMP, GPa-glucose and R-state GPa, GPb-AMP have been determined (Johnson and Barford, 1990). Glycogen phosphorylase is homodimeric (although the R state enzyme tends to be tetrameric in the absence of its glycogen substrate) consisting of two domains (N-terminal and C-terminal domains) made up of a core of pleated β-sheets flanked by α-helices. The catalytic site is located between the two domains (N-terminal domain and C-domain) and is close to the pyridoxal phosphate cofactor (attached to Lys680). Access to the catalytic site via a channel some 12 Å in length is severely restricted in diameter in one region in the T state. The catalytic site is flanked by six loops; one of the loops is the 280s loop carrying Asp283, Asn284 and Phe285, which lock access to the catalytic site when the enzyme is inhibited. Structural changes accompanying allosteric transitions in GP can be summarized as:

1. *Phosphorylative allosteric transition*: Phosphorylation transforms the T state (inactive) GPb to the R state (active) Gpa, resulting in the burial and ordering of N-terminal 16 residues, yet the exposure and disordering of C-terminal 5 residues. Salt bridge, His36···Asp838′ is broken so that the hydrogen bond, His36···pSer14 (phospho-Ser14) can be formed. The phosphate of pSer14 also forms salt bridges with Arg43 and Arg69′. The allosteric transition is primarily electrostatic.

TABLE 11.15 Allosteric effectors of muscle glycogen phosphorylase

Effector	Interaction	K_d (mM)
Allsteric activators	*Promote R conformation*	
AMP	Intersubunit AMP/ATP site	0.002
Glycogen	Storage site	1
G1P, phosphate	Active site	3
F1P, UDPG	I site	0.1–1.0
Allosteric inhibitors	*Promote T conformation*	
ATP, ADP	Intersubunit AMP/ATP site	100
Purines	I site	0.1
Glucose, G6P	Active site	5

Note: The equilibrium constants for allosteric transitions for phospho- and dephospho-GP are:

$T \underset{\text{Glucose}}{\overset{\text{AMP}}{\rightleftharpoons}} R$	L = T/R
Phospho-GP (GPa)	1/3
Dephospho-GP (GP)b	3000

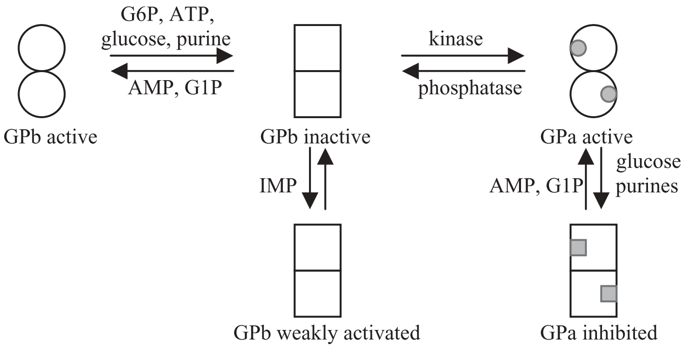


Figure 11.17 Model for allosteric transitions of glycogen phosphorylase
Glycogen phosphorylase (GP) is a homodimer consisting of two domains per protomer. The N-terminal domain (Ser1 – Gly480) includes the subunit boundary, the phosphorylation site (Ser14, green patch), the activating AMP and inhibiting G6P binding sites, the glycogen storage site and a small part of the catalytic site. The C-terminal domain (Tyr481 – Pro841) complements the catalytic site, the covalently bound pyridoxal phosphate (near the catalytic site) and inhibitory purine site (I site). The allosteric transitions between the R state (circle) and T state (square) of glycogen phosphorylase (GP) are affected by reversible phosphorylation between the active, phosphorylated a form (GPa) and the inactive dephosphorylated b form (GPb) as well as a number of effectors.

2. *Homotropic transition*: Allosteric activators bind to the R state whereas allosteric inhibitors bind to the T state (Table 11.13). The greatest structural changes occur in the region, residues 262–290, consisting of tower helices (α7 helices), α8 helices, the connecting 280s loop with Asp283-Asn284-Phe285 (the lock). The allosteric transition consists of rotations of one subunit relative to the other by 10° about an axis at the subunit boundary that is normal to the twofold symmetry axis. Allosteric effects are transmitted from the subunit boundary to the active site by the tower helices, which tilt and slide relative to each other. The angle between the two helices changes from +20° in the T state to −80° in the R state. In the T state, the tower helices pack antiparallel with a cluster of hydrogen bonds between Asn270 (Asn270′) and Asn274 (Asn274′). Tyr262 is in van der Waals contact with Pro281 of the 280s loop of the other subunit. This loop blocks access to the catalytic site in the T state. Asp283 is in indirect contact with the 5′-phosphate of the cofactor pyridoxal phosphate. In the R state, the tower helices change the angle of tilt with respect to each other. The Asn/Asn contacts are broken and there are no contact between Tyr262 and the 280s loop of the other subunit. The 280s loop is disordered and no longer blocks access to the catalytic site. These movements enable ionic residues, including a conversion of pyridoxal phosphate from monoanionic to dianionic, at the catalytic site, to adopt correct orientation to promote substrate binding and catalysis. The transition has action-at-distance/conformational effect.
3. *Heterotropic transition*: Heterotropic interactions may be transmitted to the regulatory sites by the tower helices and by changes at the subunit contacts. The simultaneous binding of allosteric activator and allosteric inhibitor gives rise to a structure intermediate between the T and the R structures. The heterotropic effect is exerted throughout the subunit interface contacts.
4. *Comparison of allosteric enzymes, GP, phosphofructokinase (PFK) and aspartate carbamoyltransferase (CT)*: Despite the diversity in the subunit–subunit interfaces, the structures have in common the feature that each interface appears to be designed

to admit two possible modes of docking in response to the transition. In general, the subunit–subunit interfaces are more extensive and more highly constrained in the T state than in the R state, except where the interface region also comprises an effector recognition site. The rotational and translational movements of the subunits with respect to one another, as the structure moves from the T to R state, are achieved by rotation about an axis perpendicular to the symmetry axis of the oligomer, to be followed in some instances by a translation such that the oligomeric symmetry is largely conserved. Interactions at the subunit interfaces and at the effector binding sites are tailored to produce a switch between the T and R states and hence allow events at one binding site to be communicated to a distant binding site over 60 Å away.

11.6 ABZYME

Immunization with an appropriate antigen has been successfully used to generate immunoglobulins (Igs) with desired characteristics, such as high affinity binding to a specific molecule or a class of molecule. Structural studies of numerous antibodies and their complexes have shown that Ig binding sites can vary extensively in size, shape and charge distribution. Antibodies share many common features with enzymes. Both possess pockets of highly specific molecular recognition. They tend to be constructed from peptide loops that link elements of secondary structures. Moreover, enzymes and antibodies exploit the same palette of amino acids and the same set of molecular interactions for their recognition tasks. They employ hydrogen bonds, ion pairs, hydrophobic contacts, and van der Waals and dispersion interactions to achieve the size, shape and charge complementarily necessary to bind ligands with high specificity and affinity. The ability to modulate interactions with ligands through conformational changes is another shared attribute.

Enzymes differ from antibodies in their ability to chemically transform the ligands they bind. The active sites of enzymes are complementary to the transition states of the reaction they promote, whereas antibody binding pockets are configured for recognition of the ground state molecule. No catalysis can take place without preferential stabilization of transition state relative to ground state. Thus it should be theoretically possible to create highly selective antibody catalysis by redirecting binding energy to the recognition of transition state structures.

Chemists have created many potent enzyme inhibitors through transition state analogy and many of these compounds have served as successful haptens for creating catalytic antibodies (Benkovic, 1992; Hilvert, 2000). Because small molecules or haptens are not immunogenic, generation of a potent immune response also requires coupling the transition state analogue to a suitable carrier proteins. In order to study kinetic properties, specificity and mechanism of individual catalysis, homogeneous and chemically well-defined monoclonal antibodies must be used. Finally, effective screening for the desired catalytically antibody must be performed. The design of transition state analogues follows the practices:

1 Proper mimicry of the shape and electrostatic properties of the transition state of the target reaction

In this strategy for example, negatively charged phosphonate esters and phosphonamidates, which are effective inhibitors of hydrolytic enzymes, resemble the anionic tetrahedral transition state arising during the hydrolysis of esters and amides. Antiphosphonate and antiphosphonamidate anitbodies have been shown to hydrolyze structurally analogous

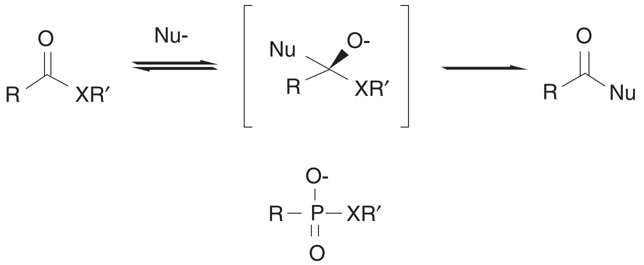


Figure 11.18 Hydrolysis of esters or amides by antiphosphonamidate Ig
The geometric and electronic properties of the tetrahedral, anionic transition state is mimicked for hydrolytic reaction by corresponding phosphonamidates used as hapten where $X{=}CH_2$, NH, O and R,R′ = alkyl or aryl groups.

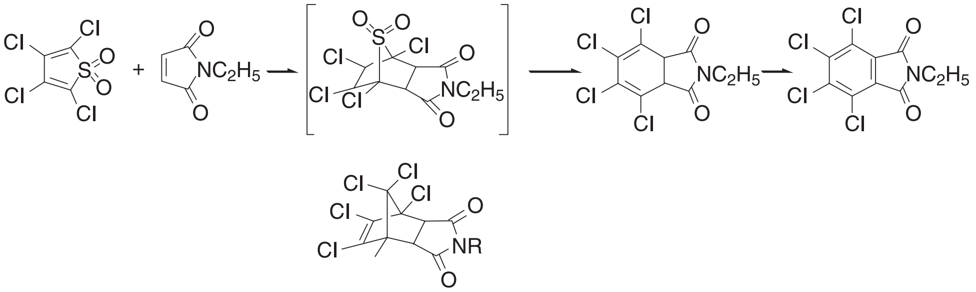


Figure 11.19 Diels-Alder reaction catalyzed by entropy trap antibody
Hexachloronorbonrnene that mimics the expected transition state is used as hapten (shown under the transition state) to elicit the antibody which effectively catalyzes the bimolecular cycloaddition.

esters and amides with rate acceleration in the range of 10^2 to 10^6 (Figure 11.18). The antibodies display classical Michaelis–Menten kinetics, exhibiting substantial substrate specificity and in some instances, high enantioselectivity.

2 Development of constrained compounds to elicit antibodies capable of catalyzing reaction with unfavorable entropies of activation

By utilizing binding energy to freeze out rotational and translational degrees of freedom necessary to reach the transition state, antibodies can act as an entropy trap. Bimolecular reactions are particularly susceptible to proximity effects and large rate accelerations can be expected simply from increasing the effective concentrations of the two substrates at the active site of an antibody. Antibodies raised against the hexachloronorbornene hapten catalyze the Diels–Alder reaction between dienes and alkenes, by increasing the effective concentration of the substrates in the Ig binding site (Figure 11.19).

3 Provision of suitable microenvironment for the transition state of the reaction of interest

Strain and distortion have also been utilized to achieve large rate accelerations with antibodies. In this strategy, antibody binding energy is used to force a substrate into a destabilizing microenvironment. For example, antibodies are raised against an *N*-methylated

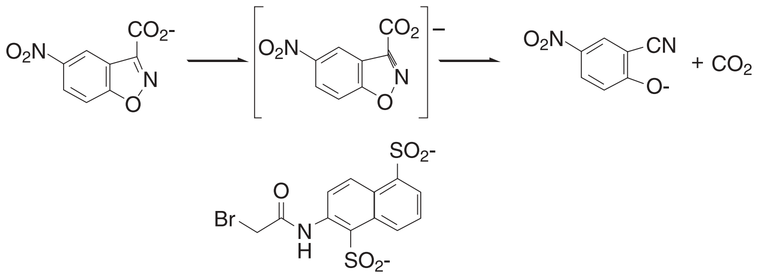


Figure 11.20 Decarboxylation catalyzed by antibody with nonpolar microenvironment
In the decarboxylation of 3-carboxybenzisoxazole by a tailored antibody, the anionic substrate is destabilized relative to the charge dispersed transition state in the nonpolar microenvironment of the antibody binding site. The hapten used to elicit the antibody is shown.

porphyrin, which has a nonplanar structure corresponding to a distorted substrate conformation, and are shown to facilitate the metalation of mesoporphyrins. Antibody catalyzed decarboxylation provides another example of this approach. The large rate accelerations observed in enzymatic decarboxylation are believed to arise, in part, from partitioning of the negatively charged carboxylate into a relatively apolar binding site where it is destabilized relative to the charge-delocalized transition state. Hydrophobic haptens have yielded antibodies with decarboxylase activity in the range of 10^4 to 10^7 over the background (Figure 11.20). In essence, the antibody provides a tailored and preorganized solvent microenvironment for the reaction.

Because many important chemical transformations, including aldol condensations, S_N2 substitutions, E2 eliminations are sensitive to solvent microenvironment, this strategy is likely to be increasingly exploited in the development of a wide variety of catalytic antibodies.

4 Introduction of catalytic groups or cofactors at the binding sites

Many enzymes promote catalyses by forming the transient enzyme intermediates involving their catalytic groups to lower the activation energies of the reactions. The difficulty in exploiting catalytic groups in antibodies is that of correct positioning within the binding pocket. Charge complementarity between hapten and antibody has been used in a number of cases to elicit acids and bases in the Ig binding sites. Catalytic functionality can also be introduced into existing antibody combining sites through semisynthesis or by site-directed mutagenesis. Enzymes often use metal ions and/or cofactors/coenzymes to promote a wider spectrum of reactions. These metal ions and low molecular weight coenzymes can be incorporated into the combining sites of antibodies to achieve the desired effect.

Monoclonal antibodies have been elicited with appropriate haptens to catalyze a variety of chemical reactions. These include hydrolyses of esters, amides and peptides (Janda *et al.*, 1989; Iverson and Lerner, 1989), stereospecific cyclization (Neppar *et al.*, 1987), Claisen rearrangement (Hilvert and Nared, 1988), redox reactions (Shokat *et al.*, 1988), and β-elimination reaction (Shokat *et al.*, 1989). In one case, an antibody raised to a phosphonate ester accelerated the hydrolysis of carboxyl esters by 6.25×10^6 (Tramontano *et al.*, 1988), which is in the range of those for known esterolytic enzymes. Metal cofactors have been successfully introduced in antibody combining sites by the judicious design of the immunizing hapten (Iverson and Lerner, 1989). Similarly, a catalytic carboxylate residue was successfully generated in an antibody combining site by exploiting the anticipated charge complementarity between immunizing hapten and elicited antibody (Shokat *et al.*, 1989). Since antibodies can be elicited against practically any structure, the possibility exists for the design and production of catalysts, i.e. catalytic antibodies, with virtually unlimited specificity. Such antibodies could have many industrial and medical applications (Schultz, 1988; Lerner and Tramontano, 1987).

Immunoglobulins possess high affinity and unmatched structural specificity toward virtually any molecule. The experimental challenge is to harness and to select from the enormous numbers and diversity of the immunological response, molecular frameworks in the form of antibodies able not only to bind a given substrate but also to catalyze a preselected chemical transformation.

Several approaches have been enlisted to improve the odds for creating catalytic antibodies, including:

1. *Cloning of the immunological response into* E. coli: Individual Fab molecules behave as a whole antibodies in terms of antigen recognition with retention of the specificity and affinity of the parent. The approach to exploring the rich diversity of the antibody system lies in the expression of the repertoires of heavy and light chains followed by their combination *in vitro*. The construction of combinatorial libraries assembled from clones of light and heavy chain fragments may capture the diversity of the immunological response while retaining the recognition and affinity properties of the parent monoclonals (Better *et al.*, 1988; Sastry *et al.*, 1989). It is an effective way to optimize the pairing of light and heavy chains for antigen recognition and possible catalysis. Two issues related to the implementation of the combinatorial libraries are: i) their rapid screening; and (ii) improved means of antibody expression.
2. *Hapten design*: Haptens are designed to introduce suitably juxtaposed catalytic residues or to modulate the hydrophilic/hydrophobic character of the antibody-binding site. This can be done by introducing desirably appropriate functionalities, a ‘bait and switch’ process (Janda *et al.*, 1990). In this approach, an additional structural element, such as charge, hydrogen bond complementarity or creation of a hydrophobic patch, is present in the immunogen but not in the substrate so that the complementary functionality induced in the antibody may assume a favorable role with the substrate molecule.
3. *Genetic or chemical modification of the antibody-combining site*: An introduction and positioning of catalytic residue(s) into antibody combining site can be achieved by:
   a) selective chemical derivatization of the combining site, permitting the incorporation of catalytic groups (Kaiser and Lawrence, 1984; Pollack and Schultz, 1989); or
   b) site-specific mutagenesis for amino acid replacement (Baldwin and Schultz, 1989; Johnson and Benkovic, 1990).

11.7 RIBOZYME

11.7.1 Characteristics of catalytic RNA

Catalytic RNAs are those that have the intrinsic ability to break and form covalent bonds and are termed ribozymes (Cech and Bass, 1986; Doherty and Doudna, 2000; DeRose,

TABLE 11.16 Some characteristics of naturally occurring RNA-acting ribozymes

Catalytic RNA	Type	Nucleophile	Reaction products
Splicing RNA:			
Group I intron	a	3′-OH of G	5′ to 3′ joined exons and intron with 5′-G and 3′-OH
Group II intron	a′	2′-OH of A	5′ to 3′ joined exons and intron with 2′-3′ lariat joined at A and 3′-OH tail
Nucleolytic RNA:			
Hammerhead	b	M^{2+} hydrate	5′-OH and 2′,3′ cyclic phosphate
Hairpin	b	M^{2+} hydrate	5′-OH and 2′,3′ cyclic phosphate
Others			
Varkud satellite	b	M^{2+} hydrate	5′-OH and 2′,3′ cyclic phosphate
Hepatitis delta virus	b	M^{2+} hydrate	5′-OH and 2′,3′ cyclic phosphate
$tRNA^{Phe}$	b	Pb^{2+} hydrate	5′-OH and 2′,3′ cyclic phosphate
Ribonucleoprotein:			
RNase P	a	H_2O	5′-phospho and 3′-OH

Note: Ribosomal ribozyme (23S rRNA) with the peptidyl transfer activity is not listed.

2002). The majority of known naturally occurring RNA catalytic activities, except ribosomal RNA, involve RNA substrates (Table 11.16). Ribozymes that act in an intramolecular reaction (*cis*-acting), catalyze only a single turnover and are usually modified during the reaction. Therefore they are considered to be acting in a quasi-catalytic manner. However, ribozymes (e.g. RNase P) can also act in *trans* in a truly catalytic manner, with a turnover greater than one and without being modified.

Eukaryotic genes are often interrupted by stretches of noncoding DNA called intervening sequences (IVS) or introns. RNA polymerases transcribe large precursor RNAs, which consists of both the exons (coding sequences) and IVS (introns). The introns are subsequently removed by a process known as RNA splicing. All ribozymes, except ribosomal RNA, perform different kinds of phosphoryl-transfer reactions, in which a transesterification reaction results in breakage of the backbone in the first step (Figure 11.21). This is brought about by the attack of a ribose hydroxyl oxygen atom on the phosphodiester bond, but is different for each type of ribozyme. The nucleophile is the adjacent 2′-hydroxyl (type b) in the nucleolytic ribozymes, whereas it is a remote 2′-hydroxyl in group II and the 3′-hydroxyl of exogenous guanosine (type a) in group I introns.

Viroids and virusoids are RNA-based pathogens. Viroids are capable of causing diseases by themselves, while virusoids are generally satellites of larger RNA viruses. During RNA replication, either of these RNAs is cleaved from a large multimeric precursor to single genome molecules by an RNA reaction. These RNAs are nucleolytic ribozymes, catalyzing the b-type cleavage generating 5′-hydroxy and 2,3′-cyclic phosphate. The cellular, mitochondrial and bacterial RNase Ps are ribonucleoproteins that carry out a hydrolytic cleavage reaction to remove the 5′-end of pre-tRNA. The RNA subunit of RNase P is catalytically active in the absence of the protein subunit. The characteristic of RNase P is specific cleavage of pre-tRNA molecules by phosphodiester hydrolysis. Cleavage is endonucleolytic, and the products of pre-tRNA cleavage retain 3′-hydroxyl and 5′-phosphate groups.

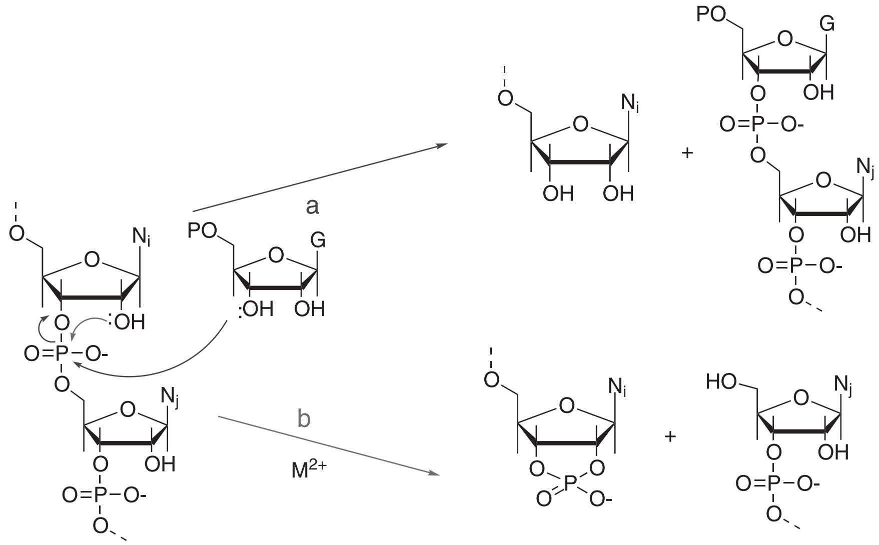


Figure 11.21 Two types of RNA cleavage catalyzed splicing and nucleolytic ribozymes
In the a-type cleavage, guanosine nucleotide (GTP/GDP/GMP) is brought into close proximity by the RNA conformation. The nucleophilic attack by the 3′-OH of guanosine at the phosphorus atom of the phosphate at the 5’ splice site results in the transesterification as the first step of guanosine meditated RNA splicing. In the b-type cleavage, the self-splicing reaction of RNA involves the divalent cation-dependent transesterification to form 2′,3′-cyclic intermediate. The cleavage using the remote 2′-OH to yield a lariat is shown in Fig 11.21.

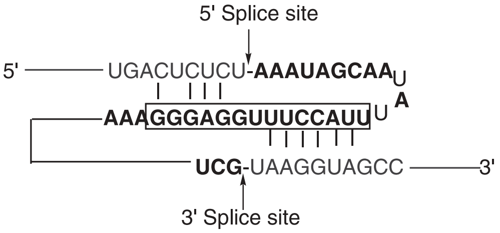


Figure 11.22 The internal guide sequence
The sequence shown is for the *Tetrahymena thermophila* pre-tRNA. Exons and introns are shown in normal and bold letters respectively. The internal guide sequence (boxed) forms base pair with both exons to form a precise alignment structure for RNA splicing. Similar structures can be drawn for most of group I intervening sequences.

11.7.2 Description of ribozymes

11.7.2.1 Group I introns. Group I Introns occur in eukaryotes and are defined based on a set of conserved sequence elements, each about ten nucleotides in length. In addition to these sequence elements, most Group I introns have a potential internal guide sequence (Figure 11.22) that aligns the 5′ and 3′ exons for splicing.

The splicing reaction catalyzed by these introns consists of two steps:

1. The RNA interacts with the guanosine nucleotide whose 3′-OH is the attacking nucleophile for the first transesterification reaction (type a cleavage), and
2. the IVS RNA interacts specifically with the stretch of pyrimidine residues at the end of the 5′ exon that is the attacking group for the second transesterification reaction.

Overall, the intron has a G added to its 5′-end and the intron is excised from the RNA precursor in the splicing reaction.

Group I introns are found in nature as self-splicing introns. For example, every copy of the nuclear 26S rRNA gene in protozoan *Tetrahymena*, is interrupted by a ~400 nucleotide IVS, which is excised in an early step of the pre-rRNA processing. The native intron is quasi-catalytic because it does not turn over. However, structurally modified forms of the intron can catalyze a variety of reactions, e.g. hydrolysis, ligation, transesterification and phosphotransfer, with multiple turnovers. The intron can be structured to recognize and transesterify exogenous substrates. The transesterification activity of a Group I intron can be used to carry out a synthetic reaction, whereby nucleotides can be joined together with the elimination of 5′-terminal G residue.

11.7.2.2 Group II introns and spliceosome. Group II introns are primarily found in fungal mitochondria and carry out a transesterification reaction, not to an external nucleotide but to the 2′-OH of an A residue in the intron. The IVS is excised as a 'lariat', a branched RNA held together by a 2′,5′ phosphodiester bond. The mechanism for splicing of mitochondrial Group II pre-mRNA is analogous to that established for nuclear pre-mRNA splicing (Padgett *et al.*, 1986). The structure of the RNA must bring the laiat-forming 2′-OH in proximity to the 5′ splice site and catalyze transesterification. The accuracy of the reaction is presumably achieved by recognition of the sequence GUGCG that occurs immediately downstream from the 5′ splice site within Group II IVS.

It is an exact analogue of the eukaryotic splicing reaction carried out by the ribonucleoprotein complex called the spliceosome. Group II introns also exhibit a-type cleavage, whereby the 5′-phosphate at the scissile bond is transferred to the splicing branch point. For the spliceosome, the RNA components adopt a base-paired structure similar to the structure of group II introns. Watson–Crick interactions position the intron G and the acceptor 2′-OH group of the lariat A residue (Figure 11.23), while the internal guide

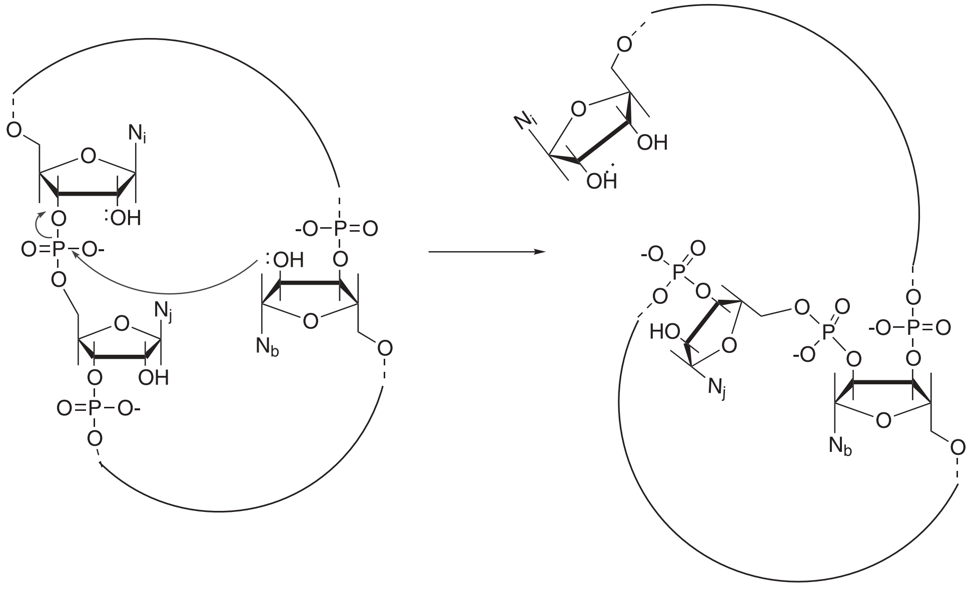


Figure 11.23 The formation of a lariat in group II intron splicing
The 5′-phosphate at the scissile bond of N_j is transferred to 2′-hydroxyl of N_b (the lariat A residue) to form a lariat structure in the initial step of splicing reaction mediated by group II intron and spliceosome.

sequence of a loop in spliceosome or group II intron binds to both the exon-intron junctions for the subsequent splicing.

The IVS of pre-mRNA in the spliceosome splicing contains three conserved sequences located at the 5′ splice site (GURAGU), at the 3′ splice site and at the site of branch formation. Deletion of a large region of an IVS outside these three conserved sequences does not generally affect splicing. This is in sharp contrast to the self-splicing IVS in which a large portion of the RNA structure is highly conserved and required for splicing activity. Thus nuclear pre-mRNA splicing requires *trans*-acting elements (present in snRNP) in addition to the conserved IVS sequences such as small nuclear ribonucleoprotein particles (snRNP), rather than *cis*-acting (the IVS itself) self-splicing. SnRNPs appear to require their protein as well as allowing their RNA components to be functional.

11.7.2.3 Hammerhead, hairpins. A number of small plant pathogenic RNAs and animal viral RNAs with less than 400 nucleotides undergo self-cleavage reaction in a site-specific manner in the presence of Mg^{2+} or other divalent cation, to produce fragments containing a 5′-hydroxyl and a 2′,3′-cyclic phosphate. The reaction proceeds by an in-line mechanism and requires a change in the conformation of the phosphate backbone resulting in inversion of configuration of the reaction product. The hammerhead RNA requires a divalent metal ion such as Mg^{2+} to mediate catalytic cleavage, i.e. the active form is an RNA-bound metal hydroxide that acts by abstracting a proton from the 2′-OH at the cleavage site.

The basic features of the hammerhead structure are the three base-paired stems I, II and III, surrounding a single-stranded central region, with the 16 conserved bases (Figure 11.24). The conserved central bases are essential for ribozyme activity. However, there appear to be few restrictions on the nonconserved nucleotides in the three base-paired

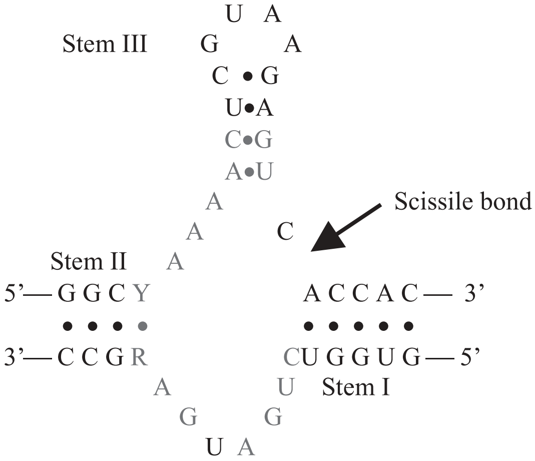


Figure 11.24 The consensus secondary structure of the hammerhead ribozyme
The secondary structure of the hammaerhead ribozyme consists of a 16–nucleotide enzyme strand and a 25-nucleotide substrate strand. Two base-paired regions (stems I and II) position the scissile bond, the third region (stem III), the hammerhead must contain enough base pairs to form a stable structure. The conserved bases (light, grayish) are required for catalytic activity and arrow indicates the cleavage site. The active site cytosine (C-17) is not conserved and can be replaced with A or U (A-17 works well but the activity of U-17 is reduced).

stems. The nucleotide 5′ to the self-cleavage site is most commonly C, occasionally A but never U or G. The bp on the inside of stem II is usually C:G, whereas the third bp of this stem is usually G:C.

The hammerhead RNA is perhaps the best characterized ribozyme, including three-dimensional structures of complexes (Pley *et al.*, 1994; Scott and Klug, 1996). The global conformation of the all hammerhead ribozymes is roughly γ-shaped fold. Stem II and stem III are approximately co-axial with stem I, with the catalytic pocket branching away from this axis. Stem II stacks directly upon stem III, forming one pseudo-continuous helix incorporating a three-strand junction where the active site cytosine is squeezed out of the helix and forced into the four-nucleotide catalytic pocket, which is formed by a sharp turn in the hammerhead enzyme strand. The phosphate backbone strands, which diverge at the three-strand junction, subsequently reunite to form stem I. Only one metal-binding site is observed near the scissile bond; Mn^{2+}/Cd^{2+} interacts with a phosphate of the conserved A at the base of stem II and with N7 of the adjacent G (of the GAR sequence). The bound metal ion is close to the site of cleavage.

The core of the hairpin ribozyme comprises two formally unpaired loops carried on adjacent arms of a four-way helical junction. The junction is not essential to the activity, but provides efficient folding to create a local structure that approximates in-line geometry for the attacking 2′-OH of the adenosine residue. In common with the structure of the hammerhead, the scissile bond, which is cleaved to a 2′,3′-cyclic phosphate, is unpaired and is bounded by helical regions.

11.7.2.4 Hepatitis delta virus RNA. Hepatitis delta virus (HDV) is a satellite virus of hepatitis B virus. The two ribozyme sequences associated with HDV are located in complementary strands of the virus. The viron referred to as the genomic strand, contains a covalently closed, single-stranded RNA of about 1700 nucleotides with about 70% self-complementary, and it can be folded into an unbranched, rod-like secondary structure typical of viroids and virusoids. The global structure of the HDV ribozyme is a double pseudoknot. A central deep cleft is formed, enclosed by two helices on one side, the pseudoknot connections and another helix on the other side. The cleavage site is found within this cleft, adjacent to the nucleobase of C75, which is juxtaposes to G1 presumably acting as general acid in the catalysis (Shih and Been, 2002).

11.7.2.5 Ribonuclease P RNA. Ribonuclease P (RNase P) is the endoribonuclease that generates the mature 5′-ends of tRNA by removal of the 5′-leader elements of pre-tRNA. Haloenzymes are composed of essential RNA and protein subunits. *E. coli* RNase P contains a protein subunit of 119 amino acids and a single RNA molecule of 377 nucleotides, with RNA exhibiting multiple turnovers *in vitro* and *in vivo*, to process tRNA precursors with extended 5′ terminal sequences (Frank and Pace, 1998; Xiao *et al.*, 2002). Although the RNA subunit of cellular RNase P contains the catalytic active site (catalyzing maturation of tRNA *in vitro*), the protein subunit is required for *in vivo* activity.

Only 21 of about 200 nucleotides in the universal core of RNase P are conserved in nucleotide identity across the phylogenic group. Most of these universally conserved nucleotides are clustered in three regions. Each RNA consists of a phylogenetically conserved core of helices and joining nucleotides that is embellished with species- or group-specific structural modifications. Sequence length variation between species is usually manifested in one of two ways:

1. by large, complex extensions to core helices; or
2. by the presence or absence of entire helices.

Whereas phylogenetically variable structures are generally dispensable for catalytic activity, they often serve to stabilize global folding of ribozymes.

RNase P cleavage reaction, which requires divalent metal ions (Mg^{2+} or Mn^{2+}), consists of three basic steps:

1. binding of the pre-tRNA substrate;
2. cleavage of the scissile phosphodiester bond; and
3. dissociation of 5′ leader and mature tRNA products.

The RNA mini-helix, consisting solely of T and acceptor stem (and including a 5′ leader), is cleaved by RNase P. RNase P occupies a patch on the substrate that overlaps the acceptor and T stems.

In addition, the 3′-CCA sequence found in all tRNAs is an important sequence-specific feature recognized by bacterial RNase P, but the 3′-CCA is not recognized by eukaryotic RNase P (Xiao *et al.*, 2002).

11.7.2.6 Ribosomal RNA. The demonstration that almost all of the protein subunits could be extracted from the ribosome, leaving most of the protein synthetic activity intact (Noller *et al.*, 1992), lends some credence to the idea that the primitive protein-synthetic activity resided in an RNA molecule. In the peptidyl transfer, the attacking nucleophile is the α-amino group of aminoacyl-tRNA, which attacks the carbon atom of the carbonyl group to yield a tetrahedral intermediate species. X-ray crystallographic study of the large ribosomal subunit indicates that the co-crystallized tetrahedral intermediate analogue is surrounded by the 23S rRNA and the nearest protein is 18 Å from the site (Nissen *et al.*, 2000). Thus the exceptional peptidyl transferase activity of ribosome is attributed to the 23S rRNA component of the macromolecular ribosome (Steitz and Moore, 2003).

11.7.3 Strategies for ribozyme catalysis

The ribozymes generally achieve a rate enhancement of ~10^5–10^6. Clearly there are differences between proteins and RNA. Proteins have a much greater diversity of functional group than RNA, including groups that are well suited for general acid–base catalysis, covalent catalysis and the formation of hydrophobic pockets. The low diversity of RNA side chains, the high charge and flexibility of its backbone, and the resultant limitations on precise positioning are also expected to limit the catalytic capabilities of RNA. However, the energetics of protein enzymes and RNA ribozymes have an underlying commonality in that both classes of biological catalysts use binding energy from interactions away from the site of bond formation and breaking to facilitate the chemical transformations.

The RNA catalysis tends to be multifactorial, with several processes contributing to an overall enhancement of reaction rate (Lilley, 2003). Three strategies adopted for ribozyme catalyses are:

11.7.3.1 General acid-base catalysis versus metal ion catalysis. The phosphotransesterification and hydrolysis reactions involve loss of a proton from the attacking hydroxyl functional group and gain of a proton on the leaving oxygen atom in the course of catalyses. Protein enzymes often use a general base for partial removal of the proton from the attacking group, and a general acid for partial addition of a proton to the leaving group to stabilize the transition state. However, RNA lacks functional groups with pK_a's near neutrality that are optimally suited for general acid–base catalysis. pK_a's near neutrality allows the highest concentration of the strongest acid or base to be present at physiological pH (Fersht, 1985). Ribozymes adopt an alternative strategy by stabilization of attacking and leaving groups with metal ions. Substantial catalysis of phosphoryl transfer can be achieved solely from well-positioned metal ions in model systems and most protein enzymes that catalyze phosphoryl transfer use metal ions.

Metal ions are also essential for folding RNA into the catalytic conformation. The charged phosphodiester backbone of RNA is coated with metal ions. For example, it has been estimated that the ~400 residue catalytic RNA from *Bacillus subtilis* RNase P binds ~100 Mg^{2+} ions under catalytic conditions (Beebe *et al.*, 1996). Proteins, by contrast, typically contain only a small number of well-defined metal binding sites. Each residue in RNA contains several potential metal ligands, such as the nonbridging phosphoryl oxygen, the 2′-hydroxyl groups and nitrogen/oxygen of the bases. The localization of metal ions to RNA in proximity to an array of ligands generates many local binding sites and opportunities to form a catalytic site, whereas metal binding sites in proteins arise from the careful placement of ligands within the context of the overall tertiary structure. The multidentate nature of metal ion coordination may make it easier for RNA to position metal ions than to position functional groups that can act as general acids and bases. In addition, the positional requirements for electrostatic stabilization of the transition state may be less stringent than for stabilization via partial proton donation or abstraction.

11.7.3.2 *Covalent catalysis.* RNA lacks the functional groups analogous to those on protein side chains that are most frequently identified as catalytic residues: the imidazole of His, the carboxylate of Asp and Glu, the alkyl amine of Lys and the sulfhydryl of Cys. These nucleophiles are either more reactive than the ultimate acceptor or capable of forming covalent intermediates in enzyme catalyses. Nevertheless, ribozymes use hydroxyls as the nucleophile effectively to carry out transphosphoesterification. The abundant presence of hydroxyl groups and their proximity to the potential cleavage sites provide an edge for the utilization of the hydroxyl nucleophile in the ribozyme catalyses. The reaction mechanism for these ribozymes is an S_N2 reaction resulting from nucleophilic attack by the 2′-O on the 3′-P, with departure of the 5′-O. The product is cyclic 2′,3′-P, and the reaction is accompanied by an inversion of configuration at the phosphate. The transition state is an oxyphosphorane intermediate with a trigonal bipyramidal structure containing the attacking 2′-O and departing 5′-O atoms at apical positions. Strictly speaking, the symmetrical oxyphosphorane might be a high-energy intermediate flanked by associative and dissociative transition states of higher and different energies.

11.7.3.3 *Nucleobase catalysis.* Nucleobases might, in principle, participate directly in catalysis. For example, the ring nitrogen N1 of adenine and N3 of cytosine could become protonated. The nucleobases might therefore participate in general acid–base chemistry, or protonated bases might stabilize negatively charged transition states electrostatically. However, the pK_a values of these bases are far from neutral (3.5 and 4.2 for adenine and cytosine respectively), whereas it is important that the pK_a is close to 7.0 for a catalyst operating near neutral pH. Therefore the pK_a must be perturbed by the local environment to be closer to 7.0 for these bases to be useful catalytically. This is possible in the highly charged context of RNA.

In hairpin ribozyme, the nucleobase of G8 is immediately adjacent to the attacking 2′-OH of A1 and could play a direct role in catalysis. The HDV ribozyme provides a well-characterized example of nucleobase catalysis. The structure of the ribozyme-product complex has been solved by crystallography (Ferré-d'Amaré *et al.*, 1998), revealing a deep cleft containing the site of cleavage. It also contains an essential cytosine base (C75 in the genomic ribozyme) that is adjacent to the 5′-O of the product. The HDV ribozyme in which C75 is replaced with uracil (C75U) is inactive, but activity could be substantially restored by the addition of a high concentration of the base imidazole, providing strong evidence for general acid–base catalysis by C75. The pH dependent rate analysis indicates that the cytosine is probably acting not as a general base but as a general acid with a pK_a of ~6.0. General base catalysis is provided by metal-bound water. The nucleobase catalysis also appears likely for the Varkud satellite ribozyme with A756 as a plausible candidate (Lilley, 2003).

11.8 REFERENCES

Abeles, R.H. and Maycock, A.I. (1976) *Account Chemical Research*, **9**, 313–19.

Ackers, G.K. and Smith, F.R. (1985) *Annual Reviews in Biochemistry*, **54**, 597–629.

Alberty, W.J. and Knowles, J.R. (1977) *Angew. Chem. Int. Ed. Engl.*, **16**, 285–93.

Arad, D., Langridge, R. and Kollman, P.A. (1990) *Journal of the American Chemistry Society*, **112**, 491–502.

Bagg, T. (2004) *Introduction to Enzyme and Coenzyme Chemistry*, Blackwell Publishers. Oxford, UK.

Baldwin, E. and Schultz, P.G. (1989) *Science*, **245**, 1104–7.

Bardsley, W.G. and Childs, R.E. (1975) *Biochemistry Journal*, **149**, 313–28.

Bardsley, W.G. and Waight, R.D. (1978) *Journal of Theoretical Biology*, **70**, 135–56.

Bartlett, G.J., Porter, C.T., Borkakoti, N. and Thornton, J.M. (2002) *Journal of Molecular Biology*, **324**, 105–21.

Beebe, J.A., Kurz, J.C. and Fierke, C.A. (1996) *Biochemistry*, **35**, 10493–505.

Bell, R.M. and Koshland, D.E. Jr. (1971) *Science*, **172**, 1253–6.

Benkovic, S.J. (1992) *Annual Reviews in Biochemistry*, **61**, 29–54.

Berridge, M.J. (1990) *Journal of Biology and Chemistry*, **265**, 9583–6.

Better, M., Chang, C.P., Robinson, R. and Horwitz, A.H. (1988) *Science*, **240**, 1041–3.

Bisswanger, H. and Schmincke-Ott, E. (eds) (1980) *Multifunctional Proteins*, John Wiley & Sons, New York.

Blow, D.M. (1976) *Account Chemical Research*, **9**, 145–52.

Blow, D.M., Birktoft, J.J. and Hartley, B.S. (1970) *Nature*, **221**, 337–40.

Bommarius, A.S. and Riebel, B.R. (2004) *Biocatalysis*, Wiley-VCH, Weinheim, Germany.

Brünger, A.T. and Karplus, M. (1991) *Acc. Chemical Research*, **24**, 54–61.

Cahn, R.S., Ingold, C.K. and Prelog, V. (1966) *Angrew. Chem. Intl Ed. Engl.*, **5**, 385–415.

Carreras, C.W. and Santi, D.V. (1995) *Annual Reviews in Biochemistry*, **64**, 721–62.

Carter, C.W. Jr. (1993) *Annual Reviews in Biochemistry*, **62**, 715–48.

Carter, P. and Wells, J.A. (1988) *Nature*, **332**, 564.

Cech, T.R. and Bass, B.L. (1986) *Annual Reviews in Biochemistry*, **55**, 599–629.

Chance, B., Pye, E.K., Ghosh, A.K. and Hess, B. (1973) *Biological and Biochemical Oscillators*, Academic Press, New York.

Chock, P.B., Rhee, S.G. and Stadtman, E.R. (1980) *Annual Reviews in Biochemistry*, **49**, 813–43.

Christianson, D.W. and Cox, J.D. (1999) *Annual Reviews in Biochemistry*, **68**, 33–57.

Cleland, W.W. (1963b) *Biochimera Biophysica Acta*, **67**, 104–72.

Cleland, W.W. (1967) *Advances in Enzymology*, **29**, 1–32.

Cleland, W.W. and Kreevoy, M.M. (1994) *Science*, **264**, 1887–90.

Cleland, W.W., Frey, P.A. and Gerlt, J.A. (1998) *Journal of Biology and Chemistry*, **273**, 25529–32.

Cohn, M. and Reed, G.H. (1982) *Annual Reviews in Biochemistry*, **51**, 365–94.

Cooper, A.A. and Stevens, T.H. (1995) *Trends in Biochemical Science*, **20**, 351–6.

Copeland, R.A. (2000) *Enzymes: A Practical Introduction to Structure, Mechanism and Data Analysis*, 2nd edn, John Wiley & Sons, New York.

Cornish-Bowden, A. (1995) *Analysis of Enzyme Kinetic Data*, Oxford University Press, Oxford, UK.

Craik, C.S. *et al.* (1987) *Journal of Cellular Biochemistry*, **33**, 199–211.

Dalziel, K. and Dickinson, F.M. (1966) *Biochemistry Journal*, **100**, 34–46.

DeRose, V.J. (2002) *Chemistry and Biology*, **9**, 961–9.

Doherty, E.A. and Doudna, J.A. (2000) *Annual Reviews in Biochemistry*, **69**, 597–615.

Dougherty, D.A. (2000) *Current Opinions in Chemistry and Biology*, **4**, 645–52.

Dutler, H. and Brandén, C.-I. (1981) *Bioorganic Chemistry*, **10**, 1–13.

Eigen, M. and Hammes, G. (1963) *Advanced Enzymology*, **25**, 1–38.

Eliot, A.C. and Kirsch, J.F. (2004) *Annual Reviews in Biochemistry*, **73**, 383–415.

Fan, F. *et al.* (1991) *Biochemistry*, **30**, 6397–401.

Ferré-d'Amaré, A.R., Zhou, K. and Doudna, J.A. (1998) *Nature*, **395**, 567–74.

Fersht, A. (1985) *Enzyme Structure and Mechanism*, 2nd edn, W.H. Freeman, New York.

Fletterick, R.J. and Sprang, S.P. (1982) *Account Chemical Research*, **15**, 361–9.

Frank, D.N. and Pace, N.R. (1998) *Annu. Rev. Biochem.* **67**, 153–80.

Frieden, C. (1993) *Trends in Biochemical Science*, **18**, 58–60.

Fujita, I., Iwasa, I. and Hansch, C. (1964) *Journal of the American Chemistry Society*, **86**, 5175–80.

Gerlt, A. (1987) *Chemistry Review*, **87**, 1079–105.

Gerlt, J.A. and Gassman, P.G. (1993) *Biochemistry*, **32**, 11943–52.

Gertler, A., Walsh, K.A. and Neurath, H. (1974) *Biochemistry*, **13**, 1302–10.

Ghisla, S. and Massey, V. (1989) *European Journal of Biochemistry*, **181**, 1–17.

Glasstone, S., Laidler, K.J. and Eyring, H.K. (1941) *The Theory of Rate Processes*, McGraw-Hill, New York.

Goldhammer, A.R. and Paradies, H.H. (1979) *Current Topics in Cell Regulation*, **15**, 109–41.

Griengl, H. (ed.) (2000) *Biocatalysis*, Springer Wien, New York.

Gutteridge, A. and Thornton, J.M. (2005) *Trends Biochom. Sci.*, **30**, 622–9.

Hadfield, A.T., Harvey, D.J., Archer, D.B. *et al.* (1994) *Journal of Molecular Biology*, **243**, 856–72.

Hahn, M.E. and Muir, T.W. (2005) *Trends in Biochemical Science*, **30**, 26–34.

Hammes, G.G. (1982) *Enzyme Catalysis and Regulation*, Academic Press, New York.

Hansch, C. and Klein, T.E. (1991) *Methods in Enzymology*, **202**, 512–43.

Hansch, C. and Hoekman, D. (1995) *Exploring QSAR: Hydrophobic, Electronic and Steric Constants*, American Chemistry Society, Washington, DC.

Hanson, K.R. and Rose, I.A. (1975) *Account Chemistry Research*, **8**, 1–10.

Hardy, L.W. and Poteete, A.R. (1991) *Biochemistry*, **30**, 9457–63.

Hashimoto, Y. *et al.* (1996) *Journal of Biochemistry*, **119**, 145–50.

Herschlag, D. (1988) *Bioorganic Chemistry*, **16**, 62–96.

Hilvert, D. (2000) *Annual Reviews in Biochemistry*, **69**, 751–93.

Hilvert, D. and Nared, K.D. (1988) *Journal of the American Chemistry Society*, **110**, 5593–4.

Huber, R.E., Gupta, M.N. and Khare, S.K. (1994) *International Journal of Biochemistry*, **26**, 309–18.

Iverson, B.L. and Lerner, R.A. (1989) *Science*, **243**, 1184–8.

Janda, K.D., Benkovic, S.J. and Lerner, R.A. (1989) *Science*, **244**, 437–40.

Janda, K.D., Weinhouse, M.I., Schloeder, D.M. *et al.* (1990) *Journal of the American Chemistry Society*, **112**, 1274.

Jeffery, C.J. (2004) *Current Opinions in Structural Biology*, **14**, 663–8.

Jencks, W.P. (1987) *Catalysis in Chemistry and Enzymology*, Dover, New York. pp 170–82.

Johnson, K.A. and Benkovic, S.J. (1990) *Enzymes*, **19**, 159–211.

Johnson, L.N. and Barford, D. (1990) *Journal of Biology and Chemistry*, **265**, 2409–12.

Kaiser, E.T. and Lawrence, D.S. (1984) *Science*, **226**, 505–11.

Kanda, T., Brewer, C.F., Okada, G. and Hehre, E.J. (1986) *Biochemistry*, **25**, 1159–65.

Käppeli, O. and Sonneleitner, B. (1986) *CRC Critrical Reviews in Biochemistry*, **4**, 299–325.

Kim, H. and Patel, M.S. (1992) *J. Biol. Chem.* **267**, 5128–32.

King, E.L. and Altman, C. (1956) *Journal of Physical Chemistry*, **60**, 1375–81.

Kollman, P. (1985) *Account Chemical Research*, **18**, 105–11.

Koshland, D.E. Jr. (1958) *Proceedings of the National Academy of Sciences, USA*, **44**, 98–104.

Koshland, D.E. Jr. (1960) *Advances in Enzymology*, **22**, 45–97.

Koshland, D.E., Némethy, G. and Filmer, D. (1966) *Biochemistry*, **5**, 365–85.

Krebs, E.G. and Beavo, J.A. (1979) *Annual Reviews in Biochemistry*, **48**, 923–59.

Kuby, S. (1991) *Enzyme Catalysis, Kinetics and Substrate Binding*, CRC Press, Boca Raton, FL.

Lazdunski, M., Pettclerc, C., Chappelet, D. and Lazdunski, C. (1971) *European Journal of Biochemistry*, **20**, 124–39.

Leatherbarrow, R.J., Fersht, A.R. and Winter, G. (1985) *Proceedings of the National Academy of Sciences, USA*, **28**, 7840.

Lerner, R.A. and Tramontano, A. (1987) *Trends in Biochemical Sciences*, **12**, 427–30.

Lerner, R.A., Benkovic, S.J. and Schultz, P.G. (1991) *Science*, **252**, 659–67.

Li, Y. and Breaker, R.R. (1999) *Current Opinions in Structural Biology*, **9**, 315–23.

Lilley, D.M.J. (2003) *Trends in Biochemical Science*, **28**, 495–501.

Lipcomb, W.N. (1983) *Annual Reviews in Biochemistry*, **52**, 17–34.

Lolis, E. and Petsko, G.A. (1990) *Annual Reviews in Biochemistry*, **59**, 597–630.

Malcolm, B.A. *et al.* (1989) *Proceedings of the National Academy of Sciences, USA*, **86**, 133–7.

Marangoni, A.G. (2003) *Enzyme Kinetics: A Modern Approach*, Wiley-Interscience, Hoboken, NJ.

Markert, C.L. (ed.) (1975) *Isozymes*, Vols 1–4, Academic Press, New York.

Matsui, H., Blanchard, J.S., Brewer, C.F. and Hehre, E.J. (1989) *Journal of Biology and Chemistry*, **264**, 8714–6.

McPherson, M.J. (1991) *Directed Mutagenesis*, IRL Press, Oxford, UK.

Means, G.E. and Feeney, R.E. (1971) *Chemical Modification of Proteins*. Holden-Day Inc.

Metzler, D.E. (1979) *Advances in Enzymology*, **50**, 1–40.

Mildvan, A.S. and Cohn, M. (1970) *Advances in Enzymology*, **33**, 1–70.

Monad, J., Wyman, J. and Changeux, J.P. (1965) *Journal of Molecular Biology*, **12**, 88–118.

Morrison, J.F. and Walsh, C.T. (1988) *Advances in Enzymology Related Areas in Molecular Biology*, **61**, 201–301.

Murphy, J.E. *et al.* (1993) *Journal of the Biology Chemistry*, **268**, 21497–500.

Myer, T. and Stryer, L. (1988) *Proceedings of the National Academy of Sciences, USA*, **85**, 5051–5.

Neppar, A.D., Benkovic, S.J., Tramontano, A. and Lerner, R.A. (1987) *Science*, **237**, 1041–3.

Nickbarg, E.B., Davenport, R.C., Petsko, G.A. and Knowles, J.R. (1988) *Biochemistry*, **27**, 5948–60.

Nissen, P., Hansen, J., Bau, N. *et al.* (2000) *Science*, **289**, 920–30.

Noller, H.F., Hoffarth, V. and Zimniak, L. (1992) *Science*, **256**, 1420–4.

Northrop, D.B. (1981) *Annual Reviews in Biochemistry*, **50**, 103–31.

Ogston, A.G. (1948) *Nature*, **162**, 963.

Padgett, R.A., Grabowski, P.J., Konarska, M.M. *et al.* (1986) *Annual Reviews in Biochemistry*, **55**, 1119–50.

Page, M.I. (1977) *Angew. Chem. Intl Ed. Engl.*, **16**, 449–59.

Pai, E.F. and Schulz, G.E. (1983) *Journal of Biology and Chemistry*, **258**, 1752–7.

Perham, R.N. (2000) *Annual Reviews in Biochemistry*, **69**, 961–1004.

Perona, J.J. and Craik, C.S. (1997) *Journal of Biology and Chemistry*, **272**, 29987–90.

Perutz, M. (1990) *Mechanisms of Cooperativity and Allosteric Regulation in Proteins*, Cambridge University Press, Cambridge, UK.

Phillips, D.C. (1967) *Proceedings of the National Academy of Sciences, USA*, **57**, 484–95.

Piszkiewicz, D. and Bruice, T.C. (1967) *Journal of the American Chemistry Society*, **89**, 6237–43.

Pley, H.W., Flaherty, K.M. and McKay, D.B. (1994) *Nature*, **372**, 68–74.

Plowman, K.M. (1972) *Enzyme Kinetics*, McGraw-Hill, New York.

Pollack, S.J. and Schultz, P.G. (1989) *Journal of the American Chemistry Society*, **111**, 1929–31.

Popják, G. and Cornforth, J.W. (1966) *Biochemistry Journal*, **101**, 553–68.

Price, N.C. and Stevens, L. (2000) *Fundamentals of Enzymology*, 3rd edn, Oxford University Press, Oxford. UK.

Profy, A.T. and Schimmel, P. (1988) *Progress in Nucleic Acid Research and Molecular Biology*, **35**, 1–26.

Radzicka, A. and Wolfenden, R. (1995) *Science*, **267**, 90–3.

Reed, L.J. (1974) *Account Chemistry Research*, **7**, 40–6.

Reuben, J. (1971) *Proceedings of the National Academy of Sciences, USA*, **68**, 563–5.

Richard, J. and Cornish-Bowden, A. (1987) *European Journal of Biochemistry*, **166**, 255–72.

Roberts, D.V. (1977) *Enzyme Kinetics*, Cambridge University Press, Cambridge, UK.

Sanz, J.M. *et al.* (1992) *Biochemistry*, **31**, 8495–9.

Sastry, L., Alting-Mees, M., Huse, W. *et al.* (1989) *Proceedings of the National Academy of Sciences, USA*, **86**, 5728–32.

Schramm, V.L. (1998) *Annual Reviews in Biochemistry*, **67**, 693–720.

Schulz, A.R. (1994) *Enzyme Kinetics*, Cambridge University Press, Cambridge.

Schultz, P.G. (1988) *Science*, **240**, 426–33.

Scott, W.G. and Klug, A. (1996) *Trends in Biochemical Sciences*, **21**, 220–4.

Scrutton, N.S., Berry, A. and Perham, R.N. (1990) *Nature*, **343**, 38–43.

Secemski, I.I. and Lienhard, G.E. (1971) *Journal of the American Chemistry Society*, **93**, 3549–50.

Segel, I.H. (1975) *Enzyme Kinetics*, John Wiley & Sons, New York.

Seydoux, F., Malhotra, O.P. and Bernhard, S.A. (1974) *CRC Critical Reviews in Biochemistry*, **2**, 227–57.

Shaw, W.V. (1987) *Biochemistry Journal*, **246**, 1–17.

Sheppard, T.L., Ordoukhanian, P. and Joyce, G.F. (2000) *Proceedings of the National Academy of Sciences, USA*, **97**, 7802–7.

Shih, I.-H. and Been, M.D. (2002) *Annual Reviews in Biochemistry*, **71**, 887–917.

Shokat, K.M., Leumann, C.J., Sugasawara, R. and Schultz, P.G. (1988) *Angrew. Chem. Int. Ed. Engl.* **27**, 1172–4.

Shokat, K.M., Leumann, C.J., Sugasawara, R. and Schultz, P.G. (1989) *Nature*, **338**, 269–71.

Sigman, D.S. and Moosner, G. (1975) *Annual Reviews in Biochemistry*, **44**, 899–931.

Sinnott, M. (ed.) (1998) *Comprehensive Biological Catalysis*. Vols 1–4, Academic Press, San Diego, CA.

Smith, M. (1985) *Annual Reviews in Genetics*, **19**, 423–62.

Soukri, A. *et al.* (1989) *Biochemistry*, **28**, 2586–92.

Srere, P.A. (1987) *Annual Reviews in Biochemistry*, **56**, 89–124.

Stadtman, E.R. (1966) *Advances in Enzymology*, **28**, 41–154.

Stadtman, E.R. and Chock, P.B. (1978) *Current Topics in Cell Regulation*, **13**, 53–95.

Steitz, T.A. and Moore, P.B. (2003) *Trends in Biochemical Science*, **28**, 411–18.

Strater, N., Lipcomb, W.N., Klabunde, T., and Krebs, B. (1996) *Angew. Chem. Intl Ed. Engl.* **35**, 2024–55.

Strynadka, N.C. and James, M.N.G. (1991) *Journal of Molecular Biology*, **220**, 401–24.

Suhnel, J. and Schowen, R.L. (1991) In *Enzyme Mechanism from Isotope Effects* (ed. P.F. Cook), CRC, Boca Raton, FL.

Swain, J.F. and Gierasch, L.M. (2006) *Cun. Opin. Stuct. Biol.* **16**, 102–6.

Tanaka, Y., Tao, W., Blanchard, J.S. and Hehre, E.J. (1994) *Journal of Biology and Chemistry*, **269**, 32306–12.

Taylor, P. (1991) *Journal of Biology and Chemistry*, **266**, 4025–8.

Tejwani, G.A. (1978) *Trends in Biochemical Science*, **3**, 30–3.

Todd, A.E., Orengo, C.A. and Thornton, J.M. (2002) *Trends in Biochemical Science*, **27**, 419–26.

Topliss, J.G. (1983) *Quantitative Structure-Activity Relationship of Drugs*, Academic Press, New York.

Tramontano, A., Ammann, A.A. and Lerner, R.A. (1988*)* *Journal of the American Chemistry Society*, **110**, 2282–6.

Tsai, C.S., Tang, J.Y. and Subbarao, S.C. (1969) *Biochemistry Journal*, **114**, 529–34.

Tsai, C.S. (1978) *Biochemistry Journal*, **173**, 483–96.

Tsai, C.S. (1997) *International Journal of Biochemistry and Cell Biology*, **29**, 3325–45.

Vallee, B.L. and Riordan, J.F. (1969) *Annual Reviews in Biochemistry*, **30**, 733–94.

Walsh, C. (1979) *Enzymatic Reaction Mechanisms*, W.H. Freeman and Comp., San Francisco, CA.

Warshel, A. (1981) *Biochemistry*, **20**, 3167–77.

Warshell, A. (1998) *Journal of Biology and Chemistry*, **273**, 27035–8.

Wilkinson, G.N. (1961) *Biochemistry Journal*, **80**, 324–32.

Wold, F. (1977) *Methods in Enzymology*, **46**, 3–14.

Wolfenden, R. (1969) *Nature*, **223**, 704–5.

Wolfenden, R. (1972) *Account Chemistry Research*, **5**, 10–18.

Xiao, S., Scott, F., Fierke, C.A. and Engelke, D.R. (2002) *Annual Reviews in Biochemistry*, **71**, 165–89.

You, K. (1982) *Methods in Enzymology*, **87**, 101–26.

World Wide Webs cited

IUBMB:	http://www.chem.qmw.ac.uk/iubmb/enzyme/
BRENDA:	http://www.brenda.uni-koeln.de
CSA:	http://ebi.ac.uk/thornton-srv/database/CSA/
EzCatDB:	http://mbs.cbrc.jp/EzCatDB/
ENZYME:	http://www.expasy.org/enzyme
Enzyme information:	http://www.chem.qmw.ac.uk/iubmb/enzyme/
IntEnz:	http://www.ebi.ac.uk/intenz
KinTekSim:	http://www.kintek-corp.com/kinteksim.htm
PDBSum:	http://www.ebi.aci.uk/thornton-srv/databases/pdbsum
PROCAT:	http://www.biochem.ucl.ac.uk/bsm/PROCAT/PROCAT.html.
QSAR:	http://mmlin1.pha.unc.edu/~jin/QSAR/
RNase P Database:	http://www.mbio.ncsu.edu/RnaseP/home.html
SCOPEC:	http://www.enzyme.com/database/scopec.php
General Enzyme Resources:	Table 11.1

CHAPTER 12

SIGNAL TRANSDUCTION AND BIODEGRADATION

12.1 CHEMICAL TRANSDUCTION: METABOLISM

Biomacromolecules participate directly or catalytically in various intracellular (e.g. chemical transduction) as well as intercellular (e.g. signal transduction) activities. Chemical transductions are collectively known as metabolism. Metabolism represents the sum of the chemical changes that convert the raw materials necessary to nourish living organisms, into energy and the chemically complex finished products of cells. Metabolism consists of a large number of enzymatic reactions organized into discrete reaction sequences/pathways (Dagley and Nicholson, 1970; Saier, 1987). The metabolic pathways that are common to all living organisms, are sometimes referred to as primary metabolism (or simply metabolism). The synthesis and degradation of biomolecules termed intermediary metabolism comprises all reactions concerned with storing and generating metabolic energy and with using that energy in biosynthesis of biochemical compounds and energy storage compounds. Energy metabolism is that part of intermediary metabolism consisting of pathways that stores or generates metabolic energy. The Boehringer Mannheim chart of metabolic pathways is available as a series of images at http://expasy.houge.ch/cgi-bin/search-biochem-index. The information and links relating to metabolic pathways can be accessed from the (Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway Database at http://www.genome.jp/kegg/pathway.html (Kanehisa *et al*., 2002) and PathDB at http://www.ncgr.org/pathdb. Numerous metabolic pathways, genetic information processes (replication, transcription, translation, sorting and degradation), membrane transport, signal transduction, cellular processes and disorder/disease processes can be retrieved from the KEGG Pathway Database.

In addition to the primary metabolic reactions, which are similar in all living organisms, a number of metabolic pathway lead to the formation of compounds peculiar to a few species or even to a single chemical race only. These reactions are summed up under the term secondary metabolism, and their products are called secondary metabolites (Herbert, 1989; Mann, 1987). Although secondary metabolites are produced by microorganisms, plants and animals, most of the substances are found in the plant kingdom. The lack of mechanisms for true excretion in higher plants may result in this unequal distribution and accumulation of the products of secondary metabolism in the vacuoles, the cell walls, or in special excretory cells or spaces of plants. The KEGG Pathway site also provides information on secondary metabolisms of terpenoids, flavonoids, alkaloids and antibiotics. Living organisms are exposed to a steadily increasing number of synthetic chemicals that are without nutritive value but nevertheless ingested, inhaled or absorbed

---

Biomacromolecules, by C. Stan Tsai
Copyright © 2007 John Wiley & Sons, Inc.

by the organisms. These compounds are foreign to the organisms and are called xenobiotics. These xenobiotics are detoxified by xenobiotic metabolizing enzymes in the processes known as xenometabolism (Gorrod *et al.*, 1988). Xenometabolism generally consists of three phases. Phase 1 oxidoreductases or hydrolases oxidize/reduce or hydrolyze the xenobiotics by introducing a reactive group for the subsequent conjugation catalyzed by phase 2 transferases. In phase 3, the hydrophilic conjugates are excreted (animals) or internally stored (plants). The information for xenometabolism can be obtained from the Biocatalysis/Biodegradation Database (UM-BBD) at http://umbbd.ahc.umn.edu/index.html (Ellis *et al.*, 2000).

The metabolism serves two fundamentally different purposes: the generation of energy and the synthesis of biological molecules. Thus metabolism can be subdivided into catabolism, *w*hich produces energy and reducing power, and anabolism, which consumes these products in the biosynthetic processes. Both catabolic and anabolic pathways occur in three stages of complexity:

- stage 1, the interconversion of biopolymers and complex lipids with monomeric building blocks;
- stage 2, the interconversion of monomeric building blocks with still simpler organic intermediates; and
- stage 3, the ultimate degradation to, or synthesis from, raw materials including CO_2, H_2O and NH_3.

Catabolism involves the oxidative degradation of complex nutrient molecules such as carbohydrates, lipids and proteins. The breakdown of these molecules by catabolism leads to the formation of simpler molecules and generates the chemical energy that is captured in the form of ATP. Because catabolism is mostly oxidative, part of the chemical energy may be conserved in the coenzyme forms, NADH and NADPH. These two nicotinamide coenzymes have very different metabolic roles. NAD^+ reduction is part of catabolism and NADPH oxidation is an important aspect of anabolism. The energy released upon oxidation of NADH is coupled to the phosphorylation of ADP to ATP. In contrast, NADPH is the source of the reducing equivalent needed for reductive biosynthetic reactions. Anabolism is a synthetic process in which the varied and complex biomolecules such as biomacromolecules are assembled from simpler precursors. The ATP and NADPH provide the energy and reducing power respectively to drive endergonic and reductive anabolic reactions. Despite their divergent roles, anabolism and catabolism are interrelated in that the products of one provide the substrates of the other. Many metabolic intermediates are shared between the two processes that occur simultaneously in the cell. However, biosynthetic and degradative pathways are rarely simple reversals of one another, even though they often begin and end with the same metabolites. The existence of separate pathways is important for energetic and regulatory reasons. The degradation of biomacromolecules will be dealt with in this chapter and their biosyntheses will be described in the next chapter.

Living organisms coordinate their activities at every level of organization through complex chemical signaling systems. Intracellular communications are maintained by the synthesis or alternation of a great variety of different substances, which are often integral components of the processes they control. For example, metabolic pathways are regulated via the feedback control of allosteric enzymes by metabolites in those pathways or by the syntheses/degradations as well as covalent modifications of biomacromolecules. Intercellular signals occur through the mediation of messenger via biomacromolecular receptors, transducers and effectors in the signal transduction. Signal transduction (Beckerman, 2005;

Dickson and Mendenhall, 2004; Gomperts *et al*., 2002) is concerned with how external influences in the presence of specific agents determine what happens on the inside of their target cells. For example, hormones are as the first messenger, usually secreted by one organ in response to an environmental demand to signal a specific response from another. Among the numerous proteins inserted in the plasma membranes of cells are receptors. These possess the sites, accessible to the extracellular milieu, that bind with specificity soluble molecules, often referred to as ligands. The binding of a just a few ligand molecules may then bring about remarkable changes within the cell as it becomes activated or triggered. A typical hormone may elicit physiological effects at concentrations as low as 10^{-10} mol/L. Signal transductions in which biomacromolecules play critical roles at different levels, are classified online at STCDB (http://bibiserv.techfak.uni-bielefeld.de/stcdb/).

12.2 ELEMENTS OF SIGNAL TRANSDUCTION

12.2.1 First messengers

The natural extracellular ligands that bind and activate receptors are called first messengers. Major classes are listed in Table 12.1.

12.2.2 Receptors

A receptor (Hulme, 1990) is a biomacromolecule in/on a cell that specifically recognizes and binds a ligand, which acts as a signal molecule, i.e. first messenger. Useful information concerning receptors and their ligands can be obtained from Receptor Database (http://impact.nihs.go.jp/RDB.html) and Relibase (http://reliase.ebi.ac.uk/). Ligand receptor interactions constitute important initial steps in various cellular processes. The ligand receptor interactions initiate various signal transduction (Clevenger, 2004; Heldin and Purton, 1996) pathways that mobilize second messengers which activate/inhibit cascades of enzymes and proteins involved in specific cellular processes (Gilman, 1987; Hepher and Gilman, 1992; Kazio *et al*., 1991). Three membrane receptor groups that mediate eukaryotic transmembrane signaling processes are:

12.2.2.1 Group 1, Single-transmembrane segment catalytic receptors. Proteins consisting of a single transmembrane segment with:

a) a globular extracellular domain, which is the ligand recognition site; and

b) an intracellular catalytic domain, which is either tyrosine kinase or guanylyl cyclase.

12.2.2.2 Group 2, Seven-transmembrane segment receptors. Integral membrane proteins consisting of seven transmembrane (7TM) helical segments with an extracellular recognition site for ligands and an intracellular recognition site for a GTP-binding protein, termed G-protein-coupled receptors (GPCRs). SEVENS (http://sevens.cbrc.jp/) is the database for 7TM GPCRs. The seven membrane-spanning segments are linked by three exposed loops on either side of the membrane, with the N-terminus projected to the outside and the C-terminus in the cytosol. These receptors comprise the largest superfamily of proteins with more than 800 members (Pierce *et al*., 2002) and are encoded by ~5% of the mammalian genome with widely varied ligands (Table 12.2).

All these GPCRs communicate with GTP-binding proteins (G-proteins) and hence with intracellular effector enzymes to generate or mobilize second messengers (such as

TABLE 12.1 First messengers

Messenger	Remark	Example
Hydrophilic hormones	Hormones are commonly released in small amounts at sites remote from the organs they target. The cell receptors must possess high affinity. Another consequence is that although a target cell may react in milliseconds to hormone binding, overall response time is in the range of seconds to hours.	Adrenaline, insulin, glucagon, gastrin, secretin, ACTH, FSH, GH, LH, TSH, PH, vasopressin and oxytocin.
Lipophilic hormones	The principle mode of action for lipophilic hormones is to penetrate cells where they interact with the intracellular receptors, which mediate their long-term effects. These in turn penetrate the nucleus to provoke specific mRNA synthesis by binding to promoter elements on DNA.	Steroid hormones (e.g. glucocorticoids, androgens, estrogens, progestins) and thyroxine.
Growth factors	Most culture cells in the culture medium containing nutrients and vitamins will not proliferate unless key stimulants known as growth factors are supplied. Apart from stimulating (or inhibiting) growth, they may initiate apoptosis, differentiation and gene expression. Their effects are generally long ranges by influencing the growth and function of neighboring cells.	EGF, PDGF, TGFβ, transferrin,
Cytokins	Several extracellular signaling proteins interact with cells of the immune or inflammatory systems by activating or modulating the proliferative properties of these cells. Cytokines that induce inflammation are pro-inflammatory mediators, whereas anti-inflammatory mediators reduce it. Chemokines are cytokines that bring about local inflammation by recruiting inflammatory cells by chemotaxis and subsequently activating them.	TNFα, interleukins, GM-CSF
Vasoactive agents.	Physical damage to tissues or damage caused by infection generates an inflammatory response. Specialized cells (mostly leukocytes) recruited to the site, act in concerted fashion to remove/sequester/dilute the cause of the injury. The process involves many interactions between cells and numerous extracellular messengers. To enable the recruitment of leukocytes and to retain a local accumulation of fluid, there is vasodilatation and a regional increase in vascular permeability by the action of vasoactive agents.	Histamine, serotonin, bradykinin, and eicosanoids (e.g. prostaglandins, thromboxanes and leukotrienes).
Neurotransmitters and neuropeptides	Neurotransmitters are released from the presynaptic cell into the tiny volume defined by the synaptic cleft. Individual neuron contains only very small quantities of the transmitter. The released transmitter then diffuses across the cleft and binds to receptors on the post-synaptic cell. The diffusion across the short distance that separate pre- and post-synaptic neurons is fast enough to allow rapid communication between nerves or between nerve and muscle at a neuromuscular junction.	acetylcholine, adrenaline noradrenaline, dopamine, serotonin, glutamate, glycine, GABA, enkephalins, substance P, angiotensin II, somatostatin

Notes: 1. Abbreviations used are ACTH, adrenocorticotropic hormone; FSH, follicle stimulating hormone; GH, growth hormone; LH, luteininzing hormone; PH, parathyroid hormone; TSH, thyroid stimulating hormone; EGF, epidermal growth factor, PDGF, platelet derived growth factor; TGF, tranforming growth factor; TNF, tumor necrosis factor; GM-CSF, granulocyte-moncyte colony stimulating factor; GABA, γ-aminobutyric acid.

2. The hormonal actions include: *Endocirne* denotes the action at a distance of hormones that may pervade the whole organism, searching out specific target tissues. *Paracrine* denotes the action of an extracellular messenger that takes effect only locally. When a substance affects the same cell from which it has been released, the activity is termed *autocrine*.

3. Eicosanoids are derived from arachidonic acid (5,8,11,14-eicosatetraenoic acid) and they operate over short distances as potent paracrine or autocrine agents, controlling many physiological and pathological cell function. In the context of inflammation, they cause plasma leakage, skin reddening and the sensation of pain.

4. Neuropeptides (amino acid sequences) are enkephalins (Tyr-Gly-Gly-Phe-Leu/Met), substance P (Arg-Pro-Lys-Pro-Gln-Gln-Phe-Phe-Gly-Leu-Met), angiotensin II (Asp-Arg-Val-Tyr-Ile-His-Pro-Phe) and somatostatin (Ala-Gly-Cys-Lys-Asn-Phe-Phe-Trp-Lys-Thr-Phe-Thr-Ser-Cys).

TABLE 12.2 Categories of 7TM receptors

Receptor properties	Ligands
Ligand binds in the core region of 7 transmembrane helices	11-cis-Retinal Acetylcholine Catecholamine Biogenic amine (histamine. Serotonin, etc) Nucleosides and nucleotides Leukotrienes, prostaglandins, prostacyclines, thromboxanes
Short peptide ligands bind partially in the core region and to the external loops	Peptide hormones (ACTH, glucagon, growth hormone, parathyroid hormone, calcitonin)
Ligands make several contacts with the N-terminus and the external loops	Hypothalamic glycoprotein releasing factors (TRH, GnRH)
Induce an extensive reorganization of an extended N-terminal segment	Metabotropic receptors for neurotransmitters (e.g. GABA, glutamate) Ca^{2+}-sensing receptors
Protease-activated receptors	Receptors for thrombin and trypsin

cAMP, Ca^{2+} etc.). Among the 7TM receptors, there are only a few highly conserved residues, confined almost entirely to the hydrophobic membrane spanning regions. The exposed segments vary in length from 7 up to 595 residues for the N-terminus, 12 to 359 residues for the C-terminus and 5 to 230 residues for the loops. The binding sites for most low molecular mass ligands such as acetylcholin, catecholamine, and the eicosanoids are located deep within the hydrophobic cores of their receptors. The binding sites for peptide hormones such as ACTH and glucagon are situated on the N-terminal segment and on the exposed loops linking the transmembrane helices. The receptors for glycoprotein hormones have been adapted by elongated N-terminal chains that extend well out into the extracellular aqueous environment.

The sites of attachment for the neurotransmitters on metabotropic receptors, and Ca^{2+} binding Ca^{2+}-sensors are situated externally, on specialized N-terminal extensions that can approach 600 residues in length. Binding of Ca^{2+} causes a conformational change in the extended extracellular domain that exposes residues, which then interact with the transmembrane core of the receptor. In this way, the extracellular domain of the receptor acts as its own auto-ligand. The Ca^{2+}-sensing receptor has to sense and then respond to very small changes in Ca^{2+} concentration against a high basal concentration (±0.025 mM in 25 mM). Obviously the affinity of the sensor has to be very low or it would be fully saturated at all times and under all conditions. The Ca^{2+}-sensor is endowed with an extracellular domain, which acts as a low affinity chelator, binding or releasing Ca^{2+} as its concentration varies within the physiological range. A monomer-dimer equilibrium could underlie the special binding properties of Ca^{2+}-sensor.

12.2.2.3 Group 3, Oligomeric ion channels. Ligand-gated ion channels consisting of associated protein subunits that contain several transmembrane segments. Typically, the ligands are neurotransmitters that open the ion channels upon binding. For example, the nicotinic receptor of the neuromuscular junction is composed of five subunits comprising α_2, β, γ and δ organized around a central axis of fivefold symmetry. The acetylcholine binds at pockets formed at the interfaces of the two α subunits with their γ and δ neighbors. Both binding sites must be occupied in order to activate the receptor. On the extracellular surface (synaptic space), the assemblage protrudes about 6 nm in the form of a large funnel. The central pore at the opening has a diameter of about 25 nm and this becomes narrower in the region where it traverses the membrane. The primary structure of the individual subunits indicates the presence of extensive stretches of sequence homology (35–40% identity). Examination of the sequence data indicates that each subunit possesses four stretches, M1, M2, M3 and M4, in which hydrophobic amino acids predominate. These hydrophobic stretches are membrane spanning sequences. Organized as α-helices, the transmembrane segments would be just sufficient in length (5 nm) to span the membrane as oily rods. The five subunits of the nicotinic receptor are each understood to traverse the membrane four times (M1–M4). The channel is lined by all α-helical M2 segments together with contributions form M1. The transmembrane segments are packed so that the channel is lined by the M2 segments. The polar surfaces of the amphipathic M2 segments in each of five subunits tend to orientate toward each other, the nonpolar surfaces seeking the nonpolar environment offered by other subunits and the membrane bilayer. The ion channel is formed at the core of the structure.

Ion-channels regulated by serotonin, γ-aminobutyric acid (GABA) and glycine are related to the nicotinic receptors. GABA and Gly regulate Cl^- channels present on the inhibitory neurons of the central nervous system (CNS). When these are opened there is a tendency for Cl^- ions to enter the cells causing membrane hyperpolarization. The various channels regulated by glutamate, the major excitatory receptor of the CNS, comprise a separate family. The subunits possess only three transmembrane segments. The equivalent M2 segment presumed to line the channel is rudimentary, enters the membrane and re-emerges on the same side (cytoplasm).

On-line prediction of GPCRs from amino acid sequences can be conducted at PRED-GPCR (http://bioinformatics.biol.uoa.gr/PRED-GPCR) and structure analysis of nuclear receptor can be performed at NRSAS (http://receptors.org/NR/servers/html/).

12.2.3 Second messengers

The first messengers act from one cell to another, and the second messenger acts within the cell (Rose and Wilkie, 2000). Activation of most membrane-associated receptors/effectors generates a diffusable intracellular signal called the second messenger. Major second messengers are listed in Table 12.3.

12.2.4 Transducers: GTP-binding proteins

The stimulation of receptors by the external signals (first messengers) results in activation of transducer proteins that mobilize chemical second messengers. In all eukaryotic organisms, a family of GTP-binding/hydrolyzing proteins called G-proteins (Gilman, 1987;

TABLE 12.3 Some second messengers and their effects

Second messenger	Effect	Source
cAMP	Protein kinase activation	Adenylyl cyclase
cGMP	Protein kinase activation, ion channel regulation	Guanylyl cyclase
Ca^{2+}	Protein kinase & Ca^{2+}-regulatory protein activation	Membrane ion channel
Inositol-1,4,5-triP	Ca^{2+}-channel activation	Phospholipase C on P-inositol
Diacylglycerol	Protein kinase C activation	Phospholipase C on P-inositol
Phosphotidic acid	Ca^{2+}-channel activation, adenylyl cyclase inhibition	Phospholipase D products
Ceramide	Protein kinase activation	Phospholipase C on sphingomyelin
Nitric oxide	Cyclase activation, smooth muscle relaxation	Nitric oxide synthase
cADP-ribose	Ca^{2+}-channel activation	cADP-ribose synthase

TABLE 12.4 Some G Proteins and their physiological effects

Protein	Ligand	Effector proteins	Effect
G_s	Epinephrine, glucagon	Adenylyl cyclase	Glycogen/fat breakdown
G_s	Antidiuretic hormone	Adenylyl cyclas	Conservation of water
G_s	Lutenizing hormone	Adenylyl cyclase	Increase estrogen, progesterone synthesis
G_i	Acetyl choline	Potassium channel	Decrease heart rate
G_i/G_o	Enkephalines, endorphins, opioids	Adenylyl cyclase, ion channel	Changes in neuron electrical activity
G_q	Angiotensin	Phospholipase C	Muscle contraction, blood pressure elevation
G_{olf}	Odorant molecules	Adenylyl cylclase	Odorant detection

Strader *et al*., 1994) plays an essential transducing role in linking cell-surface receptors to effector proteins at the plasma membrane. Major G-proteins and their effectors are listed in Table 12.4. The term G-protein is generally reserved for the class of GTP-binding proteins that interact with 7TM receptors. All of these are composed of three subunits, α, β and γ referring to the class of heterotrimeric G-proteins. This distinguishes them form the other main class of GTP-binding proteins involved in signaling. These are monomeric and related to the protein products of the ras proto-oncogenes. The basic cycle of GTP binding and GTP hydrolysis, switching them between active and inactive states is common to all of them and is coupled to many cellular functions (George and O'Dowd, 2005; Haga and Takeda, 2005). Databases for G-proteins and G-protein coupled receptors are available at gpDB (http://bioinformatics.biol.biol.uoa.gr/gpDB) and GPCRDB (http://www.gpcr.org/7tm/) respectively.

12.2.4.1 Heterotrimeric G-protein.

The structural organization of heterotrimeric G protein shows that seven regions of β structure of the β-subunit are organized like the blades of a propeller and are tightly associated with the γ-chain. Together they behave as a single entity, βγ-subunits. The α-subunit contains the nucleotide binding site and forms contacts with one face of the β-subunit. The diversity of the heterotrimeric G-proteins is principally a function of their α-subunits. The whole assembly is anchored to the membrane by hydrophobic attachments, one at the N-terminal of the α-subunit and the other at the C-terminal of the γ-subunit.

The cycle of events regulated by all GTP-binding proteins starts and ends with GDP situated in the guanine nucleotide binding site of the α-subunits. Throughout the cycle, the G-protein remains associated with the effector enzyme through its α- or βγ-subunits, but its association with the activated receptor is transient. Upon stimulation, and association of the receptor with the G-protein-effector complex, the GDP dissociates and is replaced by GTP. The selectivity of binding and hence the progress of the cycle is determined by the 10-fold excess of GTP over GDP in cells (GTP is present at about 100 μmol/L). After activation, the contact between the G-protein and the agonist-receptor complex is weakened, thus allowing the receptor to detach and seek out further inactive G-protein molecules. This provides an opportunity for signal amplification at this point. The irreversibility of the cycle is determined by cleavage of the terminal phosphate of GTP (GTPase reaction). With the restoration of GDP in the nucleotide binding site of the α-subunits, the G-protein returns to its inactive form. The activation is thus kinetically regulated, positively by the initial rate of GDP dissociation and then negatively by the rate of GTP hydrolysis. In this way, the state of activation can be approximated by the ratio of two rate constants, each of which is under independent control:

$$\text{active/inactive} = \text{on/off} = k_{dis}/k_{cat}$$

where k_{diss} is the rate constant for the dissociation of GDP and k_{cat} is the rate constant for the hydrolysis of GTP. The GTP-bound form of the α-subunit is required to activate effector enzymes such as adenylyl cyclase and phospholipase C. The duration of this interaction lasts only as long as the GTP remains intact. As soon as it is hydrolyzed to the GDP, communication ceases and the effectors revert to their inactive state.

The GTPase activity of the main classes of G-proteins is subject to regulation through the family of RGS proteins (regulator of G-protein signaling). These interact with specific α-subunit (α_s and α_i for stimulation and inhibition of adenylate cyclase respectively) to accelerate the rate of GTP hydrolysis and act as acute regulators of a wide variety of physiological processes. There are other proteins having similar activity that act to enhance the GTPase activity of the small monomeric GTP-binding proteins such as Ras. These are known as GTPase activating proteins (GAP).

The C-terminal region of α-subunits dictates the specificity of interaction with receptors. The N-terminal sequence is the site of interaction with βγ-subunits. Crystallographic studies of α_t (stimulation of cGMP phosphodiesterase) and α_i indicate that these proteins are folded as two independent domains (Coleman *et al.*, 1994). One of these, the random domain (rd), has a structure closely related to Ras and contains the guanine nucleotide binding site. The other is a compact helical domain (hd, a bundle of six helices) that is absent from Ras and other monomeric GTP-binding proteins. All the sites of interaction linking the α-subunits to other proteins have been mapped to the conserved rd domain and the N-terminal sequence. However, the sequences of the hd domains in the different α-subunits vary considerably and so it is reasonable to assume that they might be linked to function.

It is now clear that at the functional level, the heterotrimeric G-proteins behave as if they were a dimeric combination of an α-subunit and an inseparable βγ pair. βγ-Subunits have the following functions:

- They ensure the localization, effective coupling and deactivation of the α-subunits.
- They regulate the affinity of the receptors for their activating ligands.
- They reduce the tendency of GDP to dissociate from α-subunits, thus stabilizing the inactive state.
- They are required for certain α-subunits (G_i and G_o) to undergo covalent modification.
- They interact directly with some downstream effector systems.
- They regulate receptor phosphorylation by specific receptor kinase.

12.2.4.2 Ras proteins: GTPase. The Ras proteins are often referred to as proto-oncogene products because they were first discovered as the transforming products of a group of related retroviruses. The transforming genes are fusion of the viral *gag* gene and one of the Ras genes derived from rats through which the virus had been passed. Ras proteins are all single chain polypeptides, 189 amino acids in length, bound to the plasma membranes of cells by lipidic post-translational attachments at their C-termini. They all bind guanine nucleotides (GTP and GDP) and act as GTPases. Evidence for a link to human tumors came with the finding that cultured fibroblasts transfected with DNA derived from a human tumor cell line contain a mutated form of Ras. The sequences of the Ras proteins are closely related and are regarded as the archetypes of a large superfamily. All members share some sequence homology to Ras and fall into distinct groups, called Ras, Rho, Rab, Ran and Arf. Representative functions have been assigned for

TABLE 12.5. Functions of Ras-related proteins

GTPase	Regulatory function
Ras	Cell proliferation and differentiation
Rho	Actin cytoskeleton modification, cell proliferation
Rab	Intracellular vesicle trafficking
Ran	Cell cycle transitions (S and M phases)
Arf	Activation of PLD, vesicle formation

members of Ras-related proteins (Table 12.5) in vertebrates and other eukaryotic organisms. Within each subfamily, the homologies are strong. The residues 26–45 comprise the effector region (switch region) that communicates with downstream proteins. The sequence of amino acids between 97 and 108 are responsible for interaction with guanine nucleotide exchange proteins (GEFs).

Two regions of Ras, called the switch regions (switch-1 and switch-2), change their conformation when the GDP is exchanged for GTP (Milburn *et al.*, 1990). The conformational transitions in these two switch regions are coupled so that binding of GTP brings about an ordered helix→coil transition at the N-terminus closest to the nucleotide of switch-2. Subsequent changes to the $\alpha 2$ helix effectively reorganize the components of the effector-binding interface. All this is reversed when the bound GTP is hydrolyzed to GDP. The proportion of Ras in the activated form is given by the ratio k_{diss}/k_{cat}. Some of the oncogenic viral gene products (vRas) differ from the wild-type cellular forms by single point mutations, which inhibit the GTPase activity (reducing k_{cat}), and this prolongs the active state. For this reason, a higher proportion of the activated GTP-bound form exists in vRas transformed cells. Ras lies at the center of a network of interacting pathways, is activated and modulated directly or indirectly by several receptors and, in turn, influences a large number of downstream processes:

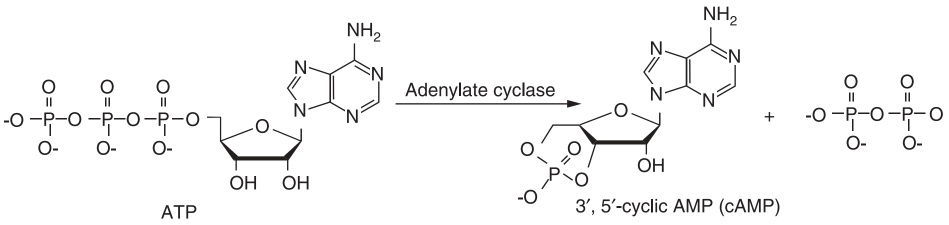
Activators:	CrKL, hSos, rasGRP, rasGRF, SmgGDS
Effectors:	Raf, RalGDS, PI-3 kinase
Inhibitors:	GAP (GTPase activating protein), neurofibromin

Among the functions directly linked to Ras are the activation of the protein kinase Raf and phosphatidylinositol 3-kinase (PI-3 kinase). Raf is the protein phosphorylating enzyme that leads to the activation of extracellular signal regulated protein kinase (ERK) and thence to the transcription of genes controlled through the serum response element (SRE).

12.3 EFFECTOR ENZYMES AND SIGNAL TRANSDUCTION

12.3.1 Adenylyl cyclase and signal transduction

Aenylyl cyclase (AC) is an effector enzyme, which converts ATP (cellular concentration of ATP = 5–10 mmol/L) to 3′,5′-cyclic adenosine monophosphate (cAMP), as



Most of the signals conveyed by AC/cAMP are mediated through phosphorylation reactions catalyzed by the protein kinases. Figure 12.1 illustrates the involvement of adenylyl cyclase as the effector enzyme in signal transduction.

Adenylyl cyclase (AC) is an integral membrane protein, generally comprising no more than 0.01–0.001% of the total membrane protein. The known physiological activators and inhibitors of AC include the subunits of the heterotrimeric G-proteins. α_s and $\beta\gamma$-subunits can activate, while α_i, α_o and $\beta\gamma$-subunits can inhibit. Also Ca^{2+} can either activate or inhibit. Type I AC reveals a number of domains. From the N-terminus, there is a predominantly hydrophobic domain, M_1, organized as six membrane-spanning α-helices linked on alternate sides of the membrane by hydrophilic loops. Next there is an extensive (40 kDa) domain, C_1, then another transmembrane domain M_2, similar to M_1, and finally at the C-terminus, a second extensive cytoplasmic domain C_2, similar to C_1 (Sunahara *et al*., 1996). Beyond the general conservation of topology, the sequences of all forms (isoforms I–VIII) of AC are similar, having 40–60% identity overall. Associated with the C_1 and C_2 domains are the subdomains C_{1a} and C_{2a} that share 50% or even higher similarity. Residues from both C_1 and C_2 contribute to ATP binding and catalysis.

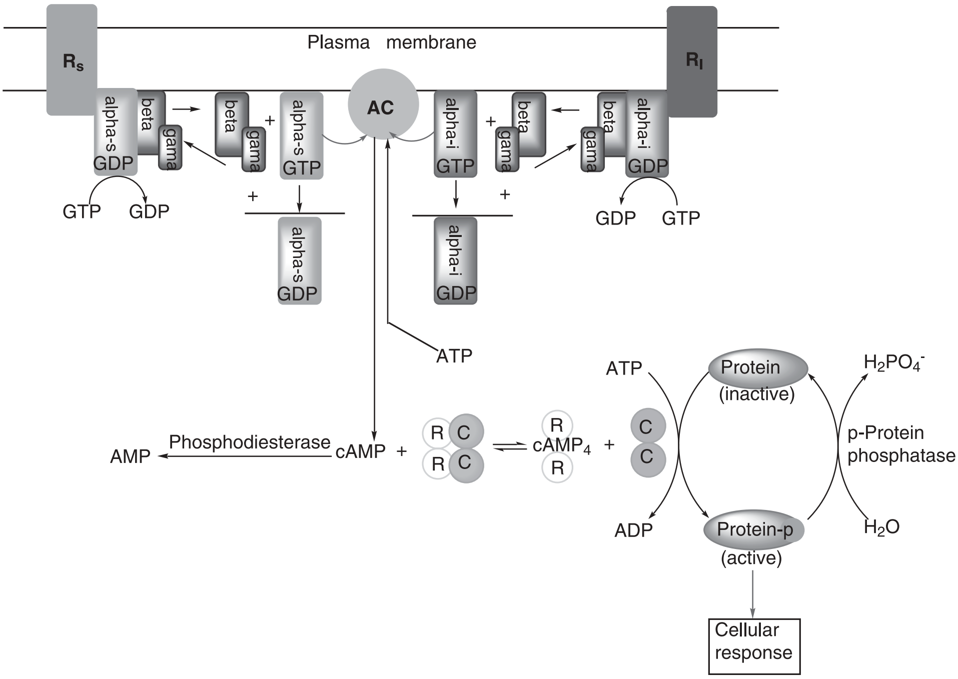


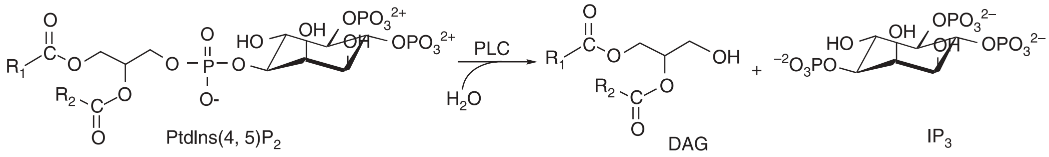
Figure 12.1 The role of adenylate cyclase in signal transduction. The binding of the first messenger to a stimulatory receptor, R_s (green) induces it to bind G_s protein which in turn stimulates the α-subunit to exchange the bound GDT for GTP. The α_s•GTP dissociates from $\beta\gamma$-subunits, until its GTPase activity hydrolyzes GTP to GDP, and activates adenylate cyclase (AC) to synthesize cAMP. The subsequent removal of cAMP *via* conversion to 5′-AMP is catalyzed by phosphodiesterases. The catalytic subunit (C_2) of cAMP-dependent protein kinase (R_2C_2), after dissociation of the regulatory dimer as R_2•$cAMP_4$, activates *via* phosphorylation of various cellular proteins leading to appropriate cellular responses. The binding of the first messenger to the inhibitory receptor, R_i (red) triggers an almost identical chain of events except that the formation of α_i•GTP inhibits AC from synthesizing cAMP. Activation (from R_s) and inhibition (from R_i) are indicated by grayish curved arrows toward AC

Organized in head-to-tail fashion and having extensive contact with each other, C_{1a} and C_{2a} are understood to comprise the catalytic center of the enzyme (Tang and Hurley, 1998). The interface between C_{1a} and C_{2a} accommodates two highly conserved nucleotide binding sites, only one of which appears to be crucial to catalysis. The binding site for α_s is distant (~3 nm) from the catalytic site and has been localized to a small region at the N-terminus of C_{1a} and to a larger region composed of a negatively charged surface and a hydrophobic groove on C_{2a}. It appears that βγ subunits potentiate the activity of type II cyclase by interacting only with the C_{2a} subdomain. This causes a conformational change that indirectly enables it to modulate the conformation of the C_{2a} subdomain and so indirectly promotes optimal alignment at the catalytic site. The inhibitory action of βγ subunits on type I cyclase must occur elsewhere since the sequences are not conserved in this region of the two subtypes.

All isoforms (I to VIII) of mammalian AC are activated by the α-subunit of G_s and all are inhibited by the so-called P-site inhibitors (analogues of adenosine such as 2′-deoxy-3-AMP). Beyond this, their responses to the many and various inhibitory ligands vary widely. They are influenced both positively and negatively by the α- and βγ-subunits of G-proteins, by Ca^{2+} and through phosphorylation by PKA and PKC. Thus isoforms of AC can be grouped into three main classes: II, IV and VII, which are activated synergistically by α_s- and βγ-subunits; V and VI, which are inhibited by α_i and Ca^{2+}; I, II and VIII, which are activated synergistically by α_s together with calmodulin dependent Ca^{2+}.

12.3.2 Phospholipase C and signal transduction

A number of second messengers are generated by lipolysis of membrane phospholipids by the action of phospholipases. Phospholipase A_2 (PLA_2) hydrolyses mostly phosphatidylcholine, yielding a lysophospholipid and releasing the fatty acid bound to the 2-position, generally arachidonate. Arachidonate is the substrate for the formation of postaglandins and leukotriens (potent signaling molecules with autocrine and paracrine effects involved in pathological processes such as inflammation). Phospholipase D (PLD) provides another, less transient, source of diacylglycerol (DAG) by cleaving phosphatidyl choline to form phosphatidate and choline. Removal of the phosphate from phosphatidate produces the DAG. The process by which inositol-1,4,5-triphosphate (IP_3) is formed in response to agonist stimulation requires the activation of a plasma-membrane associated, phosphatidylinositol specific phospholipase C (PLC), which hydrolyzes phosphatidylinositol 4,5-*bis*phosphate (PtdIns(4,5)P_2, PIP_2) such as:



The role of the plasma membrane PLC as the effector enzyme in signal transduction is shown in Figure 12.2.

A number of phophatidylinositol specific PLC isoforms have been purified and fall into three main classes, β, γ and δ (Singer *et al.*, 1997). All of the PLCs, starting from the N-terminus, possess a PH domain and then a stretch containing EF-hand motifs. All forms of PLC possess a C2 domain in the vicinity of the C-terminus. The catalytic centers, split

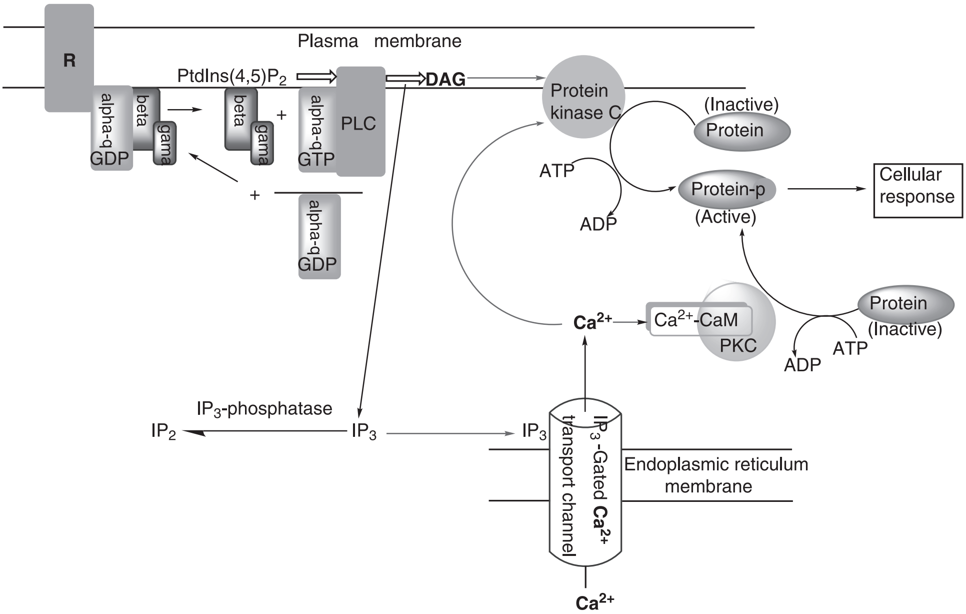

Plasma membrane
R
PtdIns(4,5)P2
DAG
alpha-q GDP
beta
gama
+
alpha-q GTP
PLC
+
alpha-q GDP
Protein kinase C
(Inactive)
Protein
ATP
ADP
Protein-p
(Active)
Cellular response
Ca2+
Ca2+-CaM
PKC
Protein
(Inactive)
ATP
ADP
IP3-phosphatase
IP2
IP3
IP3
IP3 -Gated Ca2+ transport channel
Endoplasmic reticulum membrane
Ca2+

Figure 12.2 The role of phospholipase C in signal transduction. The binding of an external signal to a membrane receptor (R) triggers GDP/GTP exchange and dissociation of the α_q•GTP complex which in turn activates phospholipase C (PLC). The hydrolysis of phosphotidyl inositol-4,5-diphosphate (PtdIns-4,5-P_2) by PLC yields two second messengers, inositol-1,4,5-triphosphate (IP_3) and diacylglycerol (DAG). The water-soluble IP_3 binds to the cytosolic side of the membrane and causes the opening of Ca^{2+} channel thus rapidly release Ca^{2+} that in turn activates numerous cellular processes through the intermediacy of calmodulin and Ca^{2+}-binding proteins. Phosphatases hydrolyze IP_3 to IP_2, IP and inositol which can be reincorporated into inositol phospholipids. The lipophilic DAG remains associated with the membrane where it activates protein kinase C (PKC) by increasing the enzyme affinity for Ca^{2+}. The activated PKC phosphorylates and activates cellular proteins leading to various cellular responses. Activation is denoted by two arrows directing toward PKC

in two separated subdomains, X and Y, possess high sequence similarity. SH2 (Src homology region 2) domains consist of about 100 residues and provide high affinity binding sites for phosphorylated tyrosine residues. The target phosphotyrosine is present in a four-residue motif with residue 1 being the phosphotyrosine and residue 4 (toward the C-terminus) usually being hydrophobic, $XY_pXX\phi$. SH3 domains consist of 55–75 amino acids that form a twisted β-barrel structure. The motif (either RXLPPLPXX or XXPPLPXR) on the target proteins, to which the SH3 domains bind, consists of a proline-rich stretch of 8–10 residues. Such sequences form extended left-handed helices having three residues per turn (called a type II polyproline helix). PH domains consist of some 100 amino acid residues characterized by a β-barrel structure. They can bind to the βγ-subunits of G-proteins and/or interact with membrane polyphosphoinositides.

The β forms of PLC are activated by a G-protein mediated mechanism involving members of the G_q, G_i or G_o families (Figure 12.2). For proteins of the G_q class, it is the α-subunit that conveys the activation signal, interacting at a site in the extended C-terminal region of these enzymes. In the case of receptors that communicate through G_i and G_o, it is the βγ heterodimer that activates PLCβ interacting at a site between the X and Y catalytic subdomains. PLCγ possesses an additional region, absent from the β and δ forms, inserted between the X and Y sequences of the catalytic domain. PLCγ is present in cells having receptors that respond to growth factors (e.g. epidermal growth factor, platelet derived growth factor and insulin receptors) and also in cells that bear receptors belonging to the immunoglobulin superfamily (e.g. T and B lymphocytes). The δ forms of PLC are of lower molecular weight and possess no domains that are unique to themselves. The regulating Ca^{2+} ions are bound at the active site of the enzyme and also to the IP_3 molecule (Essen *et al.*, 1996).

12.4 TOPICS ON SIGNAL TRANSDUCTION

12.4.1 Calcium signaling

Calcium is a more ubiquitous second messenger than cAMP, regulating a diverse range of activities, including secretion, muscle contraction, fertilization, gene transcription and cell proliferation (Putney, 2005). Calcium is well suited for the role of an intracellular messenger for two main reasons:

1. Its immediate proximity provides the opportunity for it to enter the cytosol directly, through membrane channels, either from the extracellular environment or from the internal reticular compartment.
2. Its distinctive coordination chemistry, i.e. the coordination numbers of Ca^{2+}-complexes may vary between 6 and 12, and the arrangement around the coordination sphere is flexible.

Ca^{2+} releases into the cytosolic compartment from the endoplasmic reticulum (ER) occur through two families of structurally related ion channels, inositol triphosphate receptors (IP_3Rs) and ryanodine receptors (RyRs). IP_3Rs are virtually universal, whereas RyRs are most evident in excitable cells.

Three RyR isoforms are known. Type 1 RyRs occur in skeletal muscle, where they interact directly with the voltage-sensing subunits of the Ca^{2+} channel complex. This requires the close apposition of the sarcoplasmic reticulum and the plasma membrane. In contrast, the type 2 RyRs that exist in cardiac muscle and some nerve cells, do not interact with plasma membrane Ca^{2+} channels in such a direct manner. A third isoform, type 3, is present in brain, muscle and some nonexcitable cells. RyRs are large (~5000 amino acid residues) such that they may extend into cytoplasm providing binding sites for a number of physiological modulators such as Ca^{2+}, ATP and calmodulin. IP_3Rs are also tetramers and each of the component subunits possesses six transmembrane segments. There are three isoforms of the basic subunit, type 1 to 3. For all of these, both the N- and C-termini lie within the cytosol. The N-terminal chain, which bears the IP_3 binding site, is also large (~2000 amino acids). This cytosolic domain possesses a number of modulatory sites.

Both IP_3Rs and RyRs are sensitive to local Ca^{2+} concentration, so that as cytosol Ca^{2+} increases, it induces further releases from the stores. This is known as Ca^{2+}-induced Ca^{2+}-release (CICR). As it proceeds it causes positive feedback, which this underlies the regenerative character of many Ca^{2+} signals. The opening of the IP_3R channels is modulated by IP_3, the binding of which increases the sensitivity of the channels to the surrounding Ca^{2+} concentration, so that CICR commences at resting cytosol Ca^{2+} levels. To prevent a runaway situation, which would lead ultimately to complete depletion of the stores, the Ca^{2+}-sensitivity of both channels begin to close again. The overall effect is that the dependence of channel open probability upon Ca^{2+} concentration is bell-shaped.

12.4.1.1 Calcium binding proteins. Ca^{2+} is an important second messenger in signal transduction. There are important signaling proteins that bind Ca^{2+} at regulatory sites and that can be activated by local or global increases in Ca^{2+} level. Common Ca^{2+} binding regions on proteins include the EF-hand motif. E and F denote helical region of the muscle protein parvalbumin. The Ca^{2+} binding site is formed by a loop of approximately 12 amino acids that links the two α-helices (helix-loop-helix motif). EF-hands generally occur as adjacent pairs allowing them to fold compactly. In general, the core of the EF-hand structure is reasonably conserved, but the outer regions differ with varying Ca^{2+} dissociation constants. In each motif, the chemical ligands that coordinate each Ca^{2+} ion are oxygen atoms provided by aspartate and glutamate side chains and by a peptide bond carbonyl group, and also by a water molecule. The EF-hand motifs in calmodulin occur in pairs (i.e. four EF-hands) and a single motif may bind a single Ca^{2+} ion.

Another Ca^{2+} binding region is the C2 domain, which is made up of approximately 130 amino acid residues arranging in a rigid eight-stranded, antiparallel β-sandwich. Ca^{2+} binding is confined to a region that is defined by three loops on one edge of the structure. In the C2 domain of PKCβ, five Asp residues on two of the loops contribute coordinating oxygens that help form the Ca^{2+} binding pocket, in which up to three Ca^{2+} ions can be bound in a cooperative manner. C2 domains also bind negatively charged phospholipids (e.g. phosphatidylserine). When the calcium ion enters the pocket, the negative charges are neutralized and the protein can then bind electrostatically to the anionic lipid, which may complete the coordination spheres of bound Ca^{2+}. Thus phospholipid binding requires the binding of Ca^{2+}, which acts like an electrostatic switch changing the surface potential of the protein. This allows soluble proteins with C2 domains to become membrane-associated when intracellular Ca^{2+} becomes elevated to micromolar levels. This exemplifies the recruitment of a signaling molecule through one of its domains to a specific membrane location.

The Ca^{2+}-dissociation constants (K_{dixxoc}) vary among the EF motif calcium binding proteins in which they are situated ($K_{dixxoc} = 10^{-7} - 10^{-5}$ mol/L). The Ca^{2+}-dissociation constants of C2 domains also vary widely ($K_{dixxoc} = 10^{-6}$ and 10^{-3} mol/L). Ca^{2+} can activate a wide range of Ca^{2+}-sensitive regulatory enzymes/proteins acting as effectors and transducers, as summarized in Table 12.6. The wave and oscillations of Ca^{2+} in a stimulated cell may activate specific downstream targets that respond within a particular range of Ca^{2+} concentrations or that require periodic stimulation (particular oscillation frequencies). Because of the temporal fluctuations, the target proteins must be able to sense and respond to short-term or local increases in the concentration of Ca^{2+} before it subsides back to the resting level. For signaling mechanisms in which Ca^{2+} rises and falls rapidly, it is not only the stability constant of the binding that is important, but also the rate of the 'on' and 'off' reactions.

12.4.1.2 Calmodulin and Ca^{2+}/calmodulin (CaM) dependent enzymes. Calmodulin (Adaptor?) and its isoform troponin C of skeletal muscle, represent the major Ca^{2+}-sensing proteins in animal cells. Calmodulin is a key Ca^{2+} sensor and mediator of intracellular enzyme that are not in themselves Ca^{2+}-binding proteins. Calmodulin is a highly conserved protein, present at significant levels in all eukaryotic cells. It is an acidic protein of modest size (17 kDa), consisting of a single, predominantly helical polypeptide chain with four Ca^{2+}-binding EF-hands, two at each end. The affinities of the individual Ca^{2+} binding sites are in the range of 10^{-6}–10^{-5} mol/L and adjacent sites bind Ca^{2+} with positive cooperativity so that the attachment of the first Ca^{2+} ion enhances the affinity of its neighbor. This has the effect of making the protein sensitive to small changes in the concentration of Ca^{2+} within the signaling range. Ca^{2+}-calmodulin itself has no intrinsic

TABLE 12.6 Examples of intracellular Ca^{2+}-binding proteins of vertebrate origin

Protein	Type	Domain	Probable function
Calsequestrin			Ca^{2+} buffering
Calreticulin			Ca^{2+} buffering
Parvalbumin		EF	Cytosolic Ca^{2+} buffering
Ca^{2+}-ATPase	CaM dependent		Pumps Ca^{2+} out of cell
Ca^{2+}-ATPase (sarcoplasmic)			Pumps Ca^{2+} into stores
Calmodulin		EF	Multipurpose Ca^{2+} sensing mediator
Troponin C		EF	Ca^{2+}sensing mediator of contraction
Ca^{2+}/calmodulin kinase II	CaM regulatory subunit		Multipurpose signaling
Myosin light chain kinase	Ca^{2+}/CaM dependent		Phosphorylate myosin II
Adenylyl cyclase (I, III, VIII)	Ca^{2+}/CaM dependent		Makes cyclic AMP
C-Nucleotide phosphodiesterase	Ca^{2+}/CaM dependent		Breaks down cyclic AMP
Phosphorylase b kinase	CaM regulatory subunit		Phosphorylate glycogen phosphorylase
Recoverin	Ca^{2+}-myristoyl switch	EF	Ca^{2+} sensing mediator
Calpain		EF	Protease
α-Actinin		EF	Cytoskeleton
Gelsolin			Actin severing and capping
Synaptotagmin	Putative Ca^{2+} sensor	C2	Signaling
Calcineurin	Ca^{2+}/CaM dependent	EF	Signaling, e.g. transcription
Protein kinase C (α, β1, β2, γ)		C2	Signaling
Phospholipases C		EF, C2	Signaling
Diacylglycerol kinase		EF	
Nitric oxide synthase	CaM regulatory subunit		Production of NO for signaling

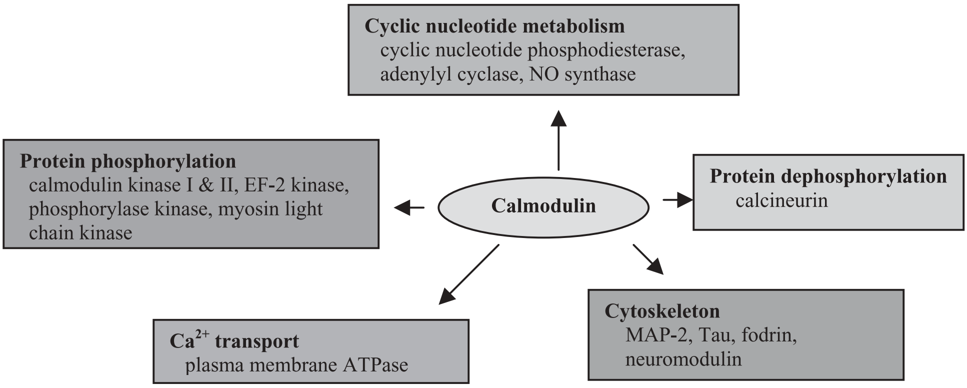

Cyclic nucleotide metabolism
cyclic nucleotide phosphodiesterase, adenylyl cyclase, NO synthase
Protein phosphorylation
calmodulin kinase I & II, EF-2 kinase, phosphorylase kinase, myosin light chain kinase
Calmodulin
Protein dephosphorylation
calcineurin
Cytoskeleton
MAP-2, Tau, fodrin, neuromodulin
Ca2+ transport
plasma membrane ATPase

Figure 12.3 Multiple signal transduction pathways initiated by calmodulin

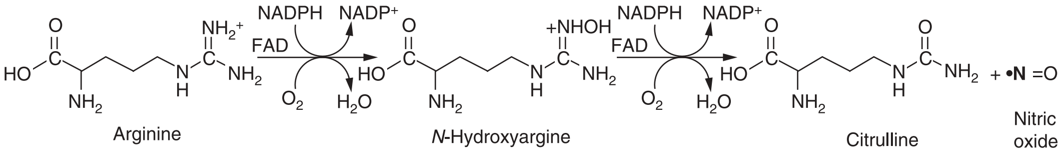
catalytic activity, its actions depending on its close association with a target enzyme/protein (Figure 12.3).

In the absence of Ca^{2+}, the central helix is shielded by the terminal helices and the protein is unable to interact with its targets. Binding of Ca^{2+} to the four sites induces a conformational change causing the terminal regions to expose hydrophobic surfaces and also exposing the central α-helical segment. Ca^{2+}-bound calmodulin binds to its targets with high affinity ($K_d \sim 10^{-9}$ mol/L). To form the bound state, the central residues of the link region unwind from their α-helical arrangement to form a hinge that allows the molecule to bend and wrap itself around the target protein. The N- and C-terminal regions approach each other and by their hydrophobic surfaces bind to it, rather like two hands holding a rope. The consequence of this interaction is a conformational change in the target protein. This state persists only as long as the Ca^{2+} concentration remains high. When it falls, the bound Ca^{2+} dissociates and calmodulin is quickly released, inactivating the target. However, an exception to this rule is CaM kinase II, which can remain in an active state after being activated by calmodulin.

Within the class of enzymes controlled by calmodulin is the Ca^{2+}/calmodulin-dependent kinase (CaM kinase). These enzymes include phosphorylase kinase, myosin light chain kinase (MLCK), which activates smooth muscle contraction, and the CaM kinases I-IV. Each of these interacts with calmodulin to convert a Ca^{2+} signal into a phosphorylation signal. The most prominent member of the family, CaM kinase II. is a broad spectrum serine/threonine kinase. It is widely expressed with particularly high levels in the brain. On activation, it undergoes autophosphorylation and this is why it remains active, even after the Ca^{2+} signal has been withdrawn. Eventually the action of phosphatases terminates the process. Calcineurin is a calmodulin-dependent Ser/Thr phosphatase. It exists as a heterodimer in mammalian cells. The catalytic subunit, calcineurin A is also familiar as protein phosphatase 2B. The regulatory subunit, calcineurin B possesses four EF-hands and binds Ca^{2+} with high affinity, but Ca^{2+}-calmodulin is still required as an additional regulatory subunit for the activation of phosphatase activity.

12.4.1.3 Calcium and nitric oxide.

Nitric oxide, which is formed via oxidation of L-arginine by the heme protein, nitric oxide synthase (NOS) acts as a neutotransmitter and as a second messenger (Bredt and Snyder, 1994):



There are three members of the NOS family: the membrane-bound endothelial enzyme called eNOS (or NOS III), a soluble enzyme in nerve and skeletal muscle called nNOS (or NOS I) and in macrophages a cytokine-inducible form called iNOS (NOS II). Both eNOS and nNOS are Ca^{2+}/calmodulin dependent enzymes and produce NO in response to a rise in cytosol Ca^{2+}. By contrast, the transcriptionally regulated iNOS binds calmodulin at resting Ca^{2+} levels and is constitutively active. It can remain activated for many hours, maximizing the cytotoxic damage that these cells can inflict. NO diffuses rapidly in solution and because it is short-lived *in vivo*, it has mostly short range (paracrine) effects. A wide range of cells are affected by NO. For example, it is a potent smooth muscle relaxant in vasculature, bronchioles, gut and genito-urinary tract. In intestinal tissue, nNOS is activated by an influx of Ca^{2+}. The NO formed diffuses into neighboring smooth muscle cells, where at nanomolar concentrations it activates a soluble guanylyl cyclase that forms cyclic 3′,5′-guanosine monophosphate (cGMP). This then activates cGMP-dependent kinase I (G kinase), which brings about relaxation by interacting with the mechanisms that regulate the cytosol Ca^{2+}, essentially keeping it low. In the cardiovascular system, eNOS in endothelial cells responds to Ca^{2+} in the similar way, to relax vascular smooth muscle and reduce blood pressure (Gyurko *et al.*, 2000). An important downstream consequence of NOS activation is the effect of NO on cellular Ca^{2+} homeostatsis. Many of the effects of NO are mediated by cGMP and the consequent activation of G kinase. This can phosphorylate and inactivate PLC and IP_3 receptors, and modulate store-operated channels (SOCs). It also inhibits Ca^{2+} release from intracellular stores in a variety of cells.

12.4.2 Phosphorylation and dephosphorylation in signaling

12.4.2.1 Protein kinases: Modulators. The importance of phosphorylation is underlined by the fact that one in three cytoplasmic proteins contains covalently bound phosphate. Phosphorylation modifies proteins by addition of negatively charged groups to serine, threonine and tyrosine, thus altering the chemical properties of proteins substantially. As a result, a protein may then recognize, bind, activate, deactivate, phosphorylate or dephosphorylate its substrate. Phosphorylation can switch enzymes on and off (subsection 11.5.2).

Protein kinases catalyze the transfer of a phosphate group from ATP to a hydroxyl residue on an amino acid side chain. There are two principal classes: serine/threonine kinases and tyrosine kinases. In both, the catalytic activity is confined to a structurally conserved domain called a protein kinase domain. The basic architecture of kinase domains is typified by the catalytic subunit of the cAMP dependent protein kinase A (PKA). The peptide chain is folded to form two lobes that are in close apposition, the cleft between them housing the catalytic site. There is also an N-terminal α-helical chain (A helix) that binds to the surface of both lobes. The lobes are connected to each other by a single polypeptide chain, which acts as a hinge (linker). The N-terminal lobe is the smaller of the two, possessing about 100 amino acids. At the N-terminus it is myristoylated. The myristoyl group occupies a pocket and provides structural stability, helping to keep the linker in contact with the large lobe. The principal feature of the small lobe is an antiparallel β-sheet. There are also two α-helical chains: the B- and C-helices. The larger lobe consists of approximately 200 residues and it possesses mostly α-helical structure, arranged around a stable four-helix bundle. Most but not all of the signals conveyed by cAMP are mediated through phosphorylation reactions catalyzed by protein kinase A (PKA).

The modulator enzyme, protein kinase A is a broad-spectrum kinase. The catalytic subunit is directed at Ser or Thr residues embedded in a sequence of amino acids, RRXS/TX (X is variable). What distinguish different forms of PKAs are the tissue-specific regulatory subunits with which they are associated. In the absence of cAMP, PKA is inactive and exists as a stable tetramer, R_2C_2, composed of two regulatory subunits (R) and two catalytic units (C) (Francis and Corbin, 1994). The tetramer is stabilized through an interaction between the catalytic sites and a pseudosubstrate sequence on the regulatory subunits, which closely resembles the substrate phosphorylation consensus sequence. For type I regulatory subunits (from skeletal muscle), it is a true pseudosubstrate motif (RRxG/Ax) in which the Ser is replaced by Gly or Ala. In type II isoforms (from cardiac muscle), the Ser is retained and the regulator is itself a substrate and undergoes phosphorylation. Autoinhibition by such motifs is a common feature of the Ser/Thr protein kinase. The binding of cAMP to the regulatory subunits is cooperative, so that the occupation of the first site of cAMP enhances the affinity of the vacant sites. The stability of the R_2C_2 complex declines by a factor in the range of 10^4–10^6 upon cAMP binding, thus liberating the catalytic units that can then phosphorylate their target proteins.

Protein kinase A also acts to regulate events in the longer term by switching on the transcription of specific genes. The phosphorylation substrate is the transcription factor cAMP-response element binding protein (CREB). As a result of phosphorylation by PKA, the CREB dimer interacts with DNA at the cAMP response element (CRE). This is an eight base-pair (bp) palindromic sequence (TGACGTCA) generally located within 100 nucleotides of the TATA box. However, transcriptional partners or co-activators are necessary for subsequent activation of gene transcription. Such factors are CREB binding protein (CBP) and p300, both large proteins that only interact with the phosphorylated form of CREB and direct the transcription factors to the transcriptional machinery situated at the TATA box (De Cesare, 1999). Over the 15–20 minutes following activation, the catalytic units of PKA diffuse into the nucleus to cause phosphorylation of CREB and the initiation of transcription. Then over 4–6 hours, transcription of the target genes gradually declines. This is probably due to dephosphorylation of CREB.

The mammalian protein kinases C (PKCs) can be subdivided into three subfamilies on the basis of sequence similarities and their mode of activation, namely cPKC, nPKC and aPKC (Table 12.7). All the PKCs have in common a catalytic domain, comprising two highly conserved subdomains, C3 (ATP binding) and C4 (substrate binding). They all contain a cystein-rich region, C1. In addition, cPKCs and nPKCs contain the C2 domains that bind Ca^{2+}. They are all characterized by the presence of a pseudosubstrate sequence that plays a role in maintaining the kinase inactive in the absence of a stimulus. Phosphorylation by the PKCs occurs only at Ser and Thr residues in the close vicinity of Arg situated in the consensus sequence, RXXS/TXRX. This is present in many proteins and as a result, PKCs can be regarded as broad-specificity protein kinases.

The deduced amino acid sequences reveal four reasonably conserved functional domains, C1–C4. Proceeding form the N-terminus, C1 and C2 constitute the regulatory domains and then C3 and C4 together constitute the catalytic domain characteristic of all kinases. C1 contains a cysteine-histidine-rich motif that coordinates two Zn atoms and this forms the binding site for diacylglycerol (DAG) and phorbol esters. The C2 domain binds negatively charged phospholipid such as phosphatidylserine and in some isoforms a Ca^{2+}-binding site responsible for the Ca^{2+} dependence of lipid binding is also present. The regulatory and the catalytic domains are linked by a hinge region. In the segment immediately

TABLE 12.7 Phenotypic changes after over-expression of PKC isoforms

Isoform	Tumorigenesis	Cell type	Change
PKCα	Tumor suppression	Rat colnic epithelial (ce) cells	Slower growth, low density
PKCα	Tumor promotion	Glioma cells	Suppression of p21
PKCβ1	Tumor promotion	Rat embryo fibroblasts	Growth in nude mice, anchorage independence, synergy with Ha-ras
PKCδ	Tumor suppression	Vascular smooth muscle cells	Slower growth, reduced expression of cyclin D and E
PKCε	Tumor promotion	Rat fibroblsts and ce cells	Growth in nude mice, anchorage independence, synergy with ras
PKCζ	Tumor suppression	v-rf transformed NIH-3T3	Reversion of transformed phenotype

Note: Based on sequence similarities and mode of activation, PKC are divided into three subfamilies:

Subfamily	Isoforms	Requirement for activation
Conventional (cPKC)	α, β1, β2, and γ	DAG, PS, Ca^{2+}
Novel (nPKC)	δ, ε, η, and θ	DAG, PS
Atypical (aPKC)	λ, τ, ζ, and μ	PS

The terminology ‘isoforms’ is better suited because the isoforms of PKC have separate characteristic substrates. Activators are; DAG, diacylglycerol and PS, phosphatidylserine.

N-terminally to the C1 domain is a stretch of amino acids that constitutes the autoinhibitory pseudosubstrate. Its sequence resembles the consensus phosphorylation site present in target proteins that are phosphorylated by PKC. However, in the pseudosubstrate the Ser is replaced by Ala. In the absence of a stimulus, the catalytic domain binds to the pseudosubstrate, causing the enzyme to fold about the hinge linking C2 and C3, resulting in a suppression of kinase activity. Activation of PKC requires phosphorylation of the catalytic domain in the activation loop and the detachment of the pseudosubstrate segment from the active site (Newton, 1997). The stimulus-mediated generation of DAG effectively plugs a hydrophilic site in the C1 domain, making the surface more hydrophobic, and this allows C1 to bury into the membrane. The DAG also reduces the Ca^{2+} requirement for the binding of the C2 domain to phospholipids hence increasing the strength of the interaction. It brings about a conformational change in the catalytic domain separating the catalytic and pseudosubstrate domains. The protein kinase is now fully active. Collectively PKC acts primarily as a modulator of the Ras signal transudation pathways that emanate from growth factor receptors. The commitment, either to promote or to suppress activity, is determined at the level of the mitogen activated protein kinases (MAP kinases). It appears that, depending on the cells and on the circumstances, different isoforms of PKC when over expressed, can either induce or suppress the formation of the transformed cell phenotypes (Table 12.7). Further Information is available at the protein kinase resource (http://www.sdsc.edu/Kinases).

12.4.2.2 *Protein tyrosine phosphatases.* Phosphoprotein phosphatases are integral components of the signaling systems operated by protein kinases (Sun and Tonks, 1994). Cloning data show the protein tyrosine phosphatases (PTPs) to be a family of multidomain proteins having exceptional diversity. They can be broadly divided into two groups, the transmembrane or receptor-like PTPs and the cytosolic PTPs. None of these are related to the serine-threonine specific phosphatases. This is in contrast to the protein kinases (Seer-Thr and Tyr specific), which share a common ancestry. Unlike the Ser-Thr phosphatases, in which substrate specificity is determined by associated targeting subunits, the Tyr phosphatases are all monomeric enzymes.

Nearly all the transmembrane PTPs are characterized by the presence of tandem repeats D1 and D2, expressing the catalytic signature motifs. However, only the membrane proximal D1 domains are catalytically active. The transmembrane PTPs are classified on the basis of their extracellular segments. They range from very short chains having no clear function, to extended structures with putative ligand binding domains, similar to those present in adhesion molecules (fibronectin repeats or immunoglobulin repeats).

The cytosolic PTPs are also classified according to their domain structures, which are understood to act as localization signals, directing the enzymes to the nucleus or cytoskeleton. Important subclasses, SHP-1 and SHP-2, possess SH2 domains. Others are characterized by the presence of PEST sequences (Pro-Glu/Asp-Sere/Thr) in the vicinity of the C-terminus. Another subclass comprises dual specific phosphatases, which can dephosphorylate at both Tyr and Ser/Thr residues. The dual specific phosphatases also have homology with Cdc25, a regulator of mitosis in *Schizosccharomyces pombe*. This activates cyclin dependent kinase-2 by dephosphorylation of adjacent threonine and tyrosine residues.

12.4.2.3 *Serine-threonine phosphatases.* Serine-threonine phosphatases are oligomeric and characterized by their association with targeting subunits. These direct them to particular locations, thus restricting their action to a limited range of substrates. The purified activities only represent the catalytic subunits, but that in the cellular environment these enzymes are coupled with targeting or inhibitory subunits. The subcellular distribution, substrate selectivity and catalytic activities are largely determined by these subunits. The serine-threonine phosphatases are classified in two superfamilies, PPP and PPM (Cohen, 1997). The PPP family consisting of PP1, PP2A, PP2B, PP4, PP5 and PP6, is characterized by the presence of three invariant amino acid motifs (–GDXHG · · · · GDXVXRG · · · · GNH–) in the catalytic domain. The PPM family comprises the Mg^{2+}-dependent PP2C and PDP (mitochondrial pyruvate dehydrogenase phosphatase).

12.4.3 Signal pathways operated by receptor protein tyrosine kinase

There are two main classes of protein tyrosine kinases (PTKs) as cell surface receptors involved in signal transduction. Phosphorylation on Tyr represents an authentic physiological process. While phosphorylation in nontransformed cells occurs mostly on Ser and Thr, phosphorylation in transformed cells includes Ser, Thr and Tyr residues. Thus Tyr phosphorylation is intimately linked to cell proliferation/transformation. Tyrosine phosphorylation is not limited to the actions of the transforming viruses or growth factors. It regulates a number of important signaling processes including:

- cell–cell and cell–matrix interactions through integrin receptors and focal adhesion sites (Giancotti, 1997)
- stimulation of the respiratory burst in phagocytic cells such as neutrophils and macrophages (Naccache *et al*., 1990)
- activation of B lymphocytes by antigen binding to the B cell receptor (Burg *et al*., 1994)
- activation of T lymphocytes by antigen presenting cells through the T cell receptor complex (Cantrell, 1996)
- the receptor for interleukin-2 (Kirken *et al*., 1993)
- high affinity receptor for immunoglobulin E (IgE) on mast cells and basophils (Li *et al*., 1992).

12.4.2.4 Tyrosine kinase-containing receptors. Tyrosine kinase containing receptors come in several different forms. They are unified by the presence of a single membrane-spanning domain and an intracellular tyrosine kinase catalytic domain. The extracellular chains vary considerably. A general feature is that ligand binding results in dimerization of the receptors. Platelet derived growth factor (PDGF) is itself a disulfide-linked dimeric ligand, which crosslinks its receptor upon binding. Epidermal growth factor (EGF), a monomeric ligand, changes the receptor conformation in the extracellular domain allowing the occupied monomers to recognize each other. For activation of all the receptor functions, the receptor molecules are brought together as dimers and orientated correctly in relation to each other so that the kinase activity of both intracellular chains may encounter target sequences of the linked receptor molecule. This enables the intermolecular cross-phosphorylation of several Tyr residues. The phosphorylated dimer becomes the active receptor. It possesses an array of phosphotyrosines, which enable it to bind proteins (adapters and enzymes) bearing SH2 domains to form receptor signaling complex. Additionally the dimerized and phosphorylated receptor has the potential of phosphorylating its targets.

The activated receptors for EGF and PDGF stimulate PLCγ. This results in the generation of DAG and IP_3, leading to the activation of PKC and a rise in the concentration of intracellular free Ca^{2+}. Furthermore, PLCγ itself becomes phosphorylated on tyrosine residues. In addition, a number of serine-threonine kinases also become activated. Most importantly, the monomeric GTPase, Ras, becomes activated. It changes from the GDP-bound state (inactive) to the GTP-bound state (active). The assembly of signaling complexes depends on the recruitment by tyrosine phosphorylated receptors of other proteins, adapters and enzymes, having SH2 or PTA domains.

The SH2 (Src homology region 2) domains are present in all nonreceptor PTKs, generally located immediately N-terminal to the kinase domain. They are also present in the adapter proteins of PTK receptors. They consist of about 100 residues and provide high affinity binding sites for phosphorylated tyrosine residues (K_d ~50–500 nmol/L). The tertiary structure consists typically of a central antiparallel β-sheet flanked on either side by two α-helices. The target phosphotyrosine is present in a four-residue motif and this binds by straddling the edge of the β-sheet. Residue 1 is the phosphotyrosine and residue 4 (toward the C-terminus) is usually hydrophobic, $XY_pXX\phi$. Each of these is held within a pocket, one on either side of the β-sheet.

Another type of domain that recognizes phosphotyrosine residues is the PTB (phosphotyrosine binding) domain (alternatively, phosphotyrosine interaction domain, PID). The specificity of interaction is determined by the sequence of amino acids immediately on the N-terminal side of the phosphorylated tyrosine, $NPXY_p$. The tertiary structure shows a β-barrel structure and a long α-helix that packs against one end. The target phosphotyrosine binds to one side of the β-barrel.

12.4.2.5 Ras signaling pathway. Ras is an important component in the signaling pathways regulating cell proliferation. The events following the activation of Ras lead to the activation of the extracellular signal regulated protein kinase (ERK, also referred to as mitogen activated protein (MAP) kinase). It enters the nucleus and is an activator of early response genes. There are two intermediate steps and both of these involve a phosphorylation. The immediate activator of ERK is MEK (MAP kinase-ERK kinase). This enzyme phosphorylates ERK on both a threonine and a tyrosine residue in the target sequence, L*T*E*Y*VATRWYRAPE. Moving further upstream, the first kinase in the cascade is Raf-1 (also called MAP kinase-kinase-kinase, MAP-KKK). The activated Ras recruits Raf-1 to the membrane and in consequence brings about kinase activation, linking ERK with the Ras pathway. Accordingly the role of Ras in the physiological situation can be regarded as that of a membrane-located recruiting sergeant. As a result of the double phosphorylation, ERK undergoes dimerization and the exposure of a signal peptide, which enables it to interact with proteins that promote its translocation into the nucleus. Within the nucleus, it catalyzes the phosphorylation of its substrate on Ser-Pro and Thr-Pro motifs. In the case of stimulation by EGF and PDGF, the activation of ERK is an absolute requirement for cell proliferation. Activation of the EGF receptor results in the rapid induction of the transcription factor c-OFS (a cytosine-inducible transcription factor). Inside the nucleus, ERK phosphorylates $p62^{TCF}$ (ternary complex factor), which then associates with $p67^{SRF}$ (serum response factor) to form an active transcription factor complex. This binds to DNA at the serum response element (SRE) in the promoter region of the c-fos gene to promote transcription of the c-fos gene.

The mitogenic signal initiated by growth factors requires an increased rate of protein synthesis and probably also the selective translation of specific mRNAs. ERK initiates ribosomal protein synthesis by regulating the binding of the initiation factor-4E (eIF-4E) to the cap of the mRNA and to the initiation factor complex. Activation of ERK results in phosphorylation of eIF-4E at Ser-209 via MNK1 and eIF-4G. Association of eIF-4E with other components of initiation complex has the effect of ironing out a hairpin loop close to the 5′ cap of the RNA and this facilitates the association of eIF-2 GTP with the small ribosomal subunit (40S) and the messenger (Sonenberg and Gingras, 1998).

12.4.4 Signaling pathways operated by nonreceptor proteins tyrosine kinase

A family of receptors are known to induce response similar to those of the receptor tyrosine kinases yet possess not intrinsic catalytic activity. Instead they recruit catalytic subunits from within the cell in the form of one or more nonreceptor protein tyrosine kinases (nrPTKs). These proteins exist within the cytosol as soluble components or they may be membrane-associated. Recruitment of nrPTKs and the consequent tyrosine phosphorylations are usually the first steps in the assembly of a substantial signaling complex consisting of a dozen or more proteins that bind and interact with each other (Hunter, 1996).

Examples of the class of receptors that recruit nrPTKs include those that mediate immune and inflammatory responses:

- The *T lymphocyte receptor* (*TCR*) is involved in detection of foreign antigens, presented together with the major histocompatibility complex (MHC). Subsequently it regulates the clonal expansion of T cells.
- *The B lymphocyte receptor* for antigen is important in the first line of defense against infection by microorganisms.
- *The interleukin-2 receptor* (*IL-2R*). The cytokine IL-2, secreted by a subset of T-helper cells, enhances the proliferation of activated T- and B-cells, increases the cytolytic activity of natural killer (NK) cells and the secretion of IgG.
- *Immunoglobulin receptors*, such as the high affinity receptor for IgE (IgER), present on mast cells and blood-borne basophils. This plays an important role in hypersensitivity and the initiation of acute inflammatory responses.
- *Erythropoietin receptors*. The cytokine erythropoietin plays an important role in the final stage of maturation of erythroid cells into mature red blood cells. The erythropoietin receptor is present on a number of cell types in addition to erthroid progenitor cells.
- *Prolactin receptors*. The pituitary hormone prolactin plays a pivotal role in the regulation of lactation. In addition it has been implicated in modulation of immune responses.

12.5 APOPTOSIS

Cell death is a normal and essential aspect of organ development. Cells dying as an integral process of development (as opposed to cells damaged by injury which undergo necrosis) commit to a form of controlled suicide called programmed cell death (PCD) or apoptosis (Brady, 2004; Jacobsen and McCarthy, 2002; Strasser *et al*., 2000). It is an essential mechanism of normal tissue homeostasis, but also play a critical role in disease state. The necrotic and apoptotic processes differ. In necrosis, the cells and their organelles swell due to the disruption of the plasma membrane. The cell contents leak out leading to inflammation and ultimately, cell disintegration. In contrast, apoptosis is characterized by loss of cell membrane phospholipid asymmetry, condensation of chromatin, reduction in nuclear size, cytoplasmic vacuolization and plasma membrane blabbing. The apoptotic cell gradually disintegrates without releasing its content. The detritus is scavenged by neighboring cells or by macrophages without the sign of inflammation. Diversified signals such as UV or γ-irradiation, oxidative damage, chemotherapeutic drugs, growth factor withdrawal, and cytokines induce apoptosis. Topics on apoptosis can be accessed at http://www.sgul.ac.uk/depts/immunology/~dash/apoptosis.

At the center of apoptosis is a family of cysteine proteases, the caspases (cysteine-aspartate proteases) that contain a Cys residue in the catalytic site and cleave their substrates at a consensus (Asp-containing) motif, Asp-Glu-Val-Asp. Caspases display a stringent requirement for Asp in the P_1 position of substrates (Nicholson and Thornberry, 1997) and can be classified into three groups based on their substrate specificity (Table 12.8).

The primary sequences of caspases show that all of the residues implicated in catalysis, in stabilization of the oxyanion intermediates, and in recognition of the P_1 Asp are conserved. The crystal structures of caspases-1 and -3 show that the large and small subunits are intimately associated with each contributing crucial residues to form a single heterodimeric catalytic domain. Furthermore, two heterodimers form a tetramer, which is the active form of the enzymes. Several characteristics of the active site are highly conserved between these two caspases:

1. The catalytic machinery involves a diad composed of the sulfhydryl group of Cys285 in close proximity to the imidazole group of His237, both from the large subunit.
2. These enzymes stabilize the oxyanion of the tetrahedral transition state through hydrogen-bond interactions with the backbone amide protons of Cys285 and Gly238.
3. Four residues (Arg179, Gln283, Arg341, Ser347), two from each subunit, appear to be involved in stabilization of the P_1 Asp of substrates.

The most significant difference between the two enzymes is the presence of a surface loop, which forms one side of the S_4 subsite in caspases-3 (residues 380–389), but absence in

TABLE 12.8 Specificity and proposed biological function for caspases

Enzyme group	P_4–P_1 specificity	Consensus	Proposed function
Group I:		WEHD	Maturation of multiple pro-inflammatory cytokine
Caspase-1 (ICE)	WEHD		
Caspase-4 (Tx)	(W/L)EHD		
Caspase-5 (Ty)	(W/L)EHD		
Group II:		DExD	Mostly effector caspases, cleavage of DxxD apoptotic substrates
CE-3	DETD		
Caspase-2	DEHD		
Caspase-3 (apopain)	DEVD		
Caspase-7	DEVD		
Group III:		(I/V/L)ExD	Mostly initiator caspases. Activation of groups II and III caspases, cleavage of non-DxxD apoptotic structures
Caspase-6 (Mch2)	VEHD		
Caspase-8 (Mch5)	LETD		
Caspase-9 (Mch6)	LEHD		
Caspase-10 (Mch4)	IEAD		
Granzyme B	IEPD		

Notes: 1. One letter codes for amino acids are used with x being any amino acids. CE-3 = caspase from *Caenorhabditis elegans*.

2. Active caspases are tetramers composed of two 17–35 kDa and two 10–12 kDa subunits. The enzymes are synthesized as zymogens (procaspases) with each precursor polypeptide giving rise to one large subunit and one small subunit during an activation process that usually involves proteolytic cleavage at Asp-X bonds.

3. Those that activate other caspases are initiator caspases and those that cleave other proteins are effector caspases. Caspases 8, 9 and 10 belong to initiator caspases whereas caspases 3, 6 and 7 belong to effector caspases. Caspase 2 is initiator/effector enzyme.

caspase-1. In caspase-1, S_4 is a large and shallow hydrophobic depression on the surface that easily accommodates a Trp side chain. In caspase-3, this subsite is significantly smaller to fit Asp, which is further stabilized by a network of hydrogen bonds.

Apoptosis is operated by two pathways (Figure 12.4). The intrinsic pathway emerges from mitochondria as a result of radiation, redox damages from detoxification and xenobiotics that cause direct DNA damage (Ravagnan *et al*., 2002). The sequences include mitochondrial release of two apoptogenic factors, cytochrome C (cyt. C) and apoptosis inducing factor (AIF). Cyt. C interacts with Apaf (apoptotic protease activating factor)-1 and procaspase-9 to form a large protein complex called apoptosome (Jiang and Wang, 2004). As the Apaf-1 is activated, a protease in turn activates procaspase-9 and begins the cascade activation of downstream caspases leading to apoptosis. The extrinsic pathway is triggered by ligand binding to specific death receptor on the cell membrane. Typical ligand-

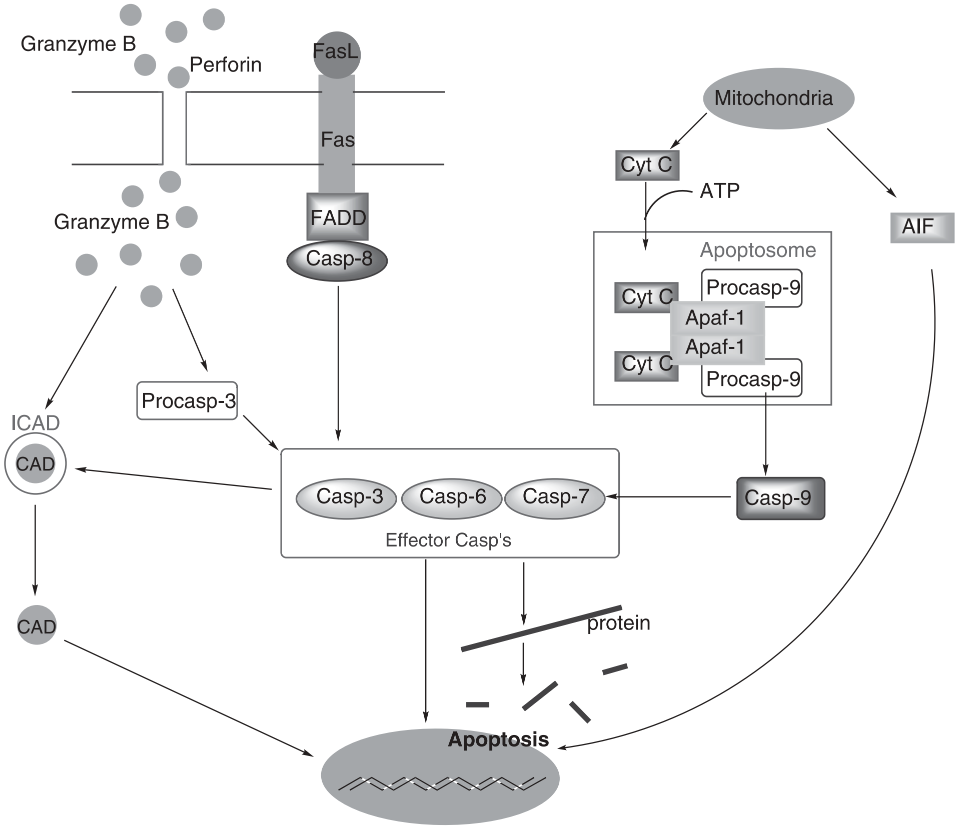


Figure 12.4 Schematic representation of caspase-mediated apoptosis. Simplified caspase-mediated apoptotic pathways, intrinsic (*via* mitochondria permeation) and extrinsic (Fas and granzyme systems) are shown. In the intrinsic pathway, the apotosome activates casp-9 which in turn activates effector caspases (casp-3, -6, and -7) leading to apoptosis. In the extrinsic pathway, the Fas system forms a signal complex consisting of FasL, Fas, FADD and procasp-8. The autocatalytically activated casp-8 then activates casp-3 leading to apoptosis. In the other extrinsic pathway triggered by the released granzyme B leads to apoptosis *via* two routes; casp-3 and CAD. Following abbreviation used are: AIF, apoptosis inducing factor; Apaf, apoptotic protease activating factor; CAD, caspase-activated DNase; casp, caspase; ICAD, inhibitor of CAD; Fas, a member of the TNF/NGF family of receptors; FasL, ligand of membrane-attached receptor Fas; FADD, death domain; NGF, nerve growth factor; TNF, tumor necrosis factor

receptor pair is the Fas-FasL system of the tumor necrosis receptor (TNR)/nerve growth factor receptor (NGFR) superfamily. The binding of the ligand stimulates oligomerization (trimerization of Fas) of the receptors via their intracellular death domain (DD). The receptors then recruit and bind via DD to the adopter proteins, FADD, which in turn associates with procaspase-8 via the death effector domain (DED) to form a signaling complex. The autocatalytic maturation of caspase-8 activates effector caspase-3 leading to apoptosis. Cytotoxic T cells release granzyme B and perforin, which form transmembrane pores and activate procaspases to initiate apoptosis (Ashkenazi *et al.*, 1998). Alternately, granzyme B may mediate caspase-independent apoptosis by activating caspase-activated DNase (CAD) via cleaving/removing CAD-associated inhibitor (ICAD). CAD is then able to interact with components such as topoisomerase II to condense chromatin leading to apoptotic DNA fragmentation.

Pro-caspases are activated by cleavage at the same consensus site, either by themselves or by other caspases. Those that activate other caspases, including themselves, are initiator caspases (caspases-8, -9, and -10) and those that cleave other protein substrates are effector caspases (caspases-3, -6 and -7). The likely roles of effector caspases in apoptosis are:

- Inactivation and destruction of inhibitors of apoptosis.
- Inactivation of the inhibitory protein (ICAD) of the caspase-activated DNase (CAD), thus allowing CAD to degrade DNA.
- Destruction of the machinery that repairs and replicates DNA.
- Proteolytic cleavages of a number of proteins that are involved in maintaining normal cellular function and of several structural proteins.
- Destruction of cellular compartments and signal transduction pathways.

Apoptosis is regulated both positively and negatively by numerous apoptogenic factors such as a family of Bcl-2 proteins (Cory and Adams, 2002). Anti-apoptotic members of the Bcl-2 family such as Bcl-2, Bcl-x and Bcl-w all contain the conserved Bcl-2 homology (BH) regions BH1, BH2 and BH3, which form a hydrophobic groove that binds to BH3 domain of other family members. Proapoptotic members of the Bcl-2 family have two types, the Bax and Bak subfamily and the BH3-only proapoptotic members such as Bid, Bad and Bim/Bod. The presence of Bax or Bak is required for many forms of apoptosis, and each type of cell needs at least one of the anti-apoptotic Bcl-2 family members to survive. Antiapoptotic Bcl-2 family members inhibit cyt.C release, possibly by modulating the ability of proapoptotic family members to facilitate opening of the voltage-dependent anion channel in the outer mitochondrial membrane. The BH3-only proteins are activated in response to various apoptotic stimuli and are able to induce apoptosis through interacting directly with Bax and Bak or binding to the hydrophobic groove and neutralizing the effect of anti-apoptoptic members such as Bcl-2 or Bcl-x. In addition to being controlled through transcription, phosphorylation and proteolytic cleavage, the proteins of Bcl-2 family are also regulated by ubiquitination and proteasome degradation systems (Yang and Yu, 2003).

12.6 HYDROLYSIS VERSUS PHOSPHOROLYSIS OF GLYCANS

Polysaccharides are degradated via hydrolysis in most cases or phosphorolysis in special cases. Glycans are hydrolyzed by either endoglycanases (endoglycosidases), which cleave the interior glycosidic linkages or exoglycanases (exoglycosidases), which remove the terminal glycose units in a stepwise manner. For example, the branched α-1,4-glucan amylose is cleaved at random internal linkages by the endoglycosidase α-amylase, producing oligosaccharide fragments. On the other hand, the exoglucosidase, β-amylase successively removes glycosyl-α-1,4-glucosyl (maltosyl) units from the nonreducing end of the glycan to yield β-maltose. Additional specifications of glycosidases are:

1. their preference for α- or β-configuration at the scissile C-1 linkage of glycans; and
2. either one of the two elementary mechanisms identified by the stereochemical outcome of the reaction, i.e., inversion or retention (Figure 12.5).

Glycosidases (EC3.2.1.–) are enzymes that are responsible for the transfer of glycosyl moieties from a donor sugar to a water acceptor with the result being hydrolysis. The catalyzed reaction is a substitution at a chiral acetal or ketal center and can occur with either of two stereochemical outcomes at the sugar anomeric center, i.e. inversion or retention. Structures and mechanisms of glycosidases have been reviewed (Henrissat and Davies, 1997; McCarter and Wither, 1994; Sinnott, 1990). The catalytic mechanism of a β-retaining glycosidase (lysozyme) has been presented (Chapter 11). Glycosidases have been classified into sequence-related families (Henrissat and Bairoch, 1993) with useful links to the databases at http://www.expasy.ch/cgi-bin/list?glycosid.txt. Enzymes within a family adopt the same structural fold and carry out their reaction with the same stereo-

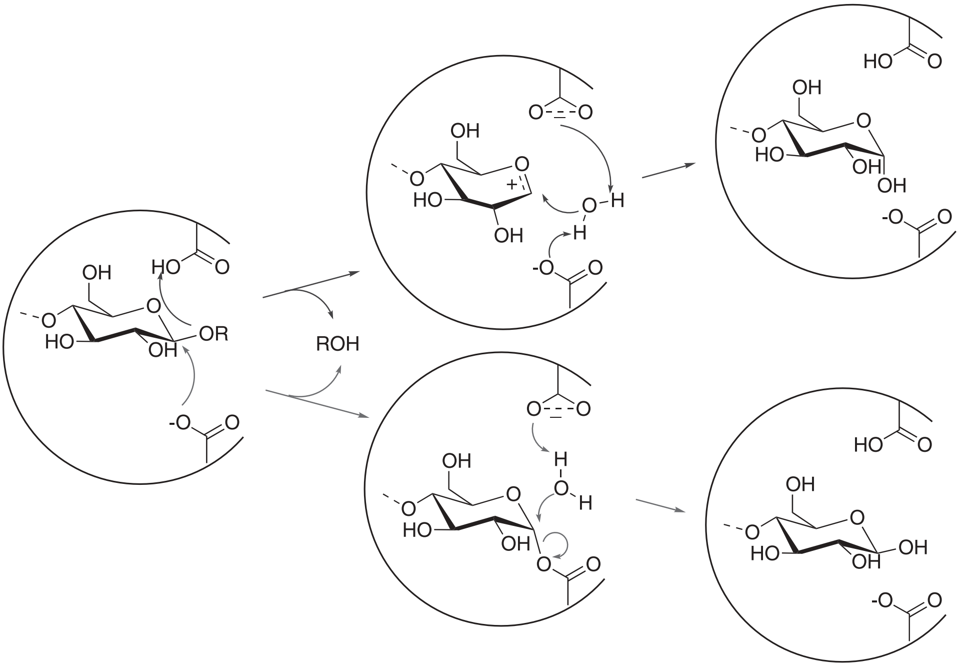


Figure 12.5 The two possible mechanisms for glycosidases. The active site of glycosidases usually consists of one or two catalytic Asp/Glu residue(s). The proposed mechanism employing two carboxyl groups, one as general acid and other as the base is exemplified for the hydrolysis of β-(1→4)-glucosidic linkages. The inverting glycosidase proceeds *via* an oxocarbonium ion intermediate (upper) and the retaining glycosidase forms a glycosyl-enzyme intermediate (lower) by the double-displacement mechanism. Both mechanisms are likely to proceed *via* oxocarbonium ion transition state in which partial charges are delocalized through the oxocarbonium ion to the two catalytic carboxyl groups. R is another glucose unit for endoglucanase, and aglycone for glucosidase

chemical outcome. Structural diversity is exhibited in their structures ranging from essentially completely α-helical to almost completely β-sheet. Many of the families are structurally (and mechanistically) related, allowing the construction of so-called clans (Jenkin *et al*., 1995).

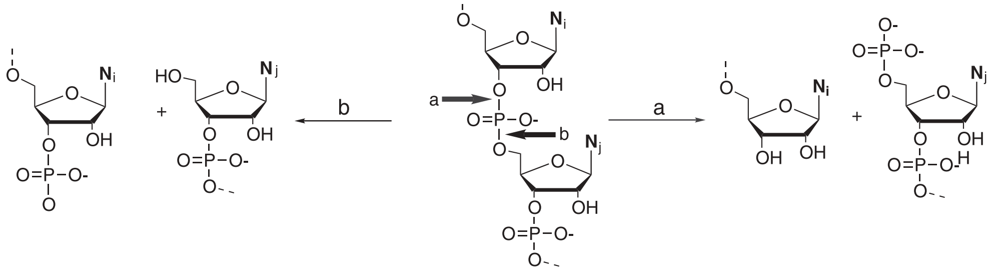
There are considerable similarities between the active sites of two classes of enzyme, consistent with their stabilization of fundamentally similar transition states (Ly and Withers, 1999). Active sites of both enzyme classes contain a pair of carboxylic acids that play key roles in the mechanisms. However, the two carboxylic acids are separated by different distances in two cases, approximately 0.95 nm for inverting glycosidases and 0.55 nm for retainers. The inverting glycosidases use a direct displacement mechanism in which the two carboxylic acids at the active site are suitably positioned such that one provides general-base catalytic assistance to the attack of water, while the other provides general-acid catalytic assistance to cleavage of the glycosidic bond. Reaction proceeds through an oxocarbonium ion-like transition state. The separation of approximately 0.95 nm is presumably just right to allow the water and substrate to bind simultaneously.

Catalysis by retaining glycosidases proceeds via a two-step double displacement mechanism involving the formation and hydrolysis of a covalent glycosyl-enzyme intermediate (both steps again proceeding through oxocarbonium ion-like transition state). The two active-site carboxylic acids have somewhat different roles in this case. One acts as the nucleophile, attacking at the sugar anomeric center to form the glycosyl-enzyme intermediate. The other carboxyl group functions as an acid/base catalyst, protonating the glycosidic oxygen in the first step (general-acid catalysis) and deprotonating the water in the second step (general-base catalysis). The shorter distance between these two residue (~0.55 nm) is therefore consistent with the need for direct attack of the nucleophile.

The typical enzyme for catalyzing phosphorolysis of glycan is glycogen phosphorylase (Fletterick and Madsen, 1980), which degrades glycogen to D-glucose-1-phosphate (G1P). Organic phosphate is an acceptor for the glucosyl transfer. The enzyme will release a glucose unit that is at least five units from a branch point at which glycogen branching enzyme is required to hydrolyzes α(1→6)-linked glycosyl units. This allows the glycogen phosphorylase reaction to go to completion. The enzyme is subjected to regulations by covalent modification and a wide variety of allosteric effectors (Chapter 11), and requires coenzyme, pyridoxal phosphate for the catalysis as depicted in Figure 12.6.

12.7 NUCLEOLYSIS OF NUCLEIC ACIDS

The hydrolytic degradation of nucleic acids is catalyzed by nucleases, which are phosphodiesterases because they cleave the phosphodiesteric bonds hydrolytically (Linn *et al*., 1993). The cleavage can potentially occur on either side of the phosphorus atom, i.e. on the 3′-side (a) and the 5′-side (b):



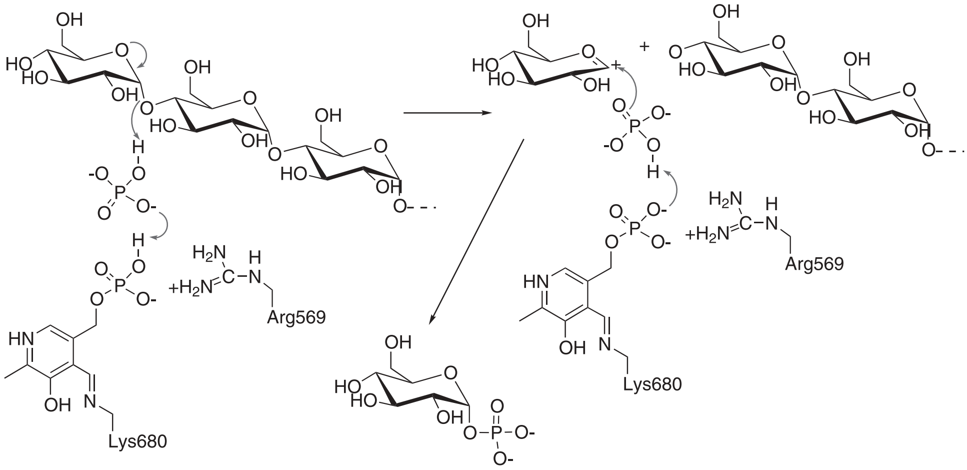


Figure 12.6 Possible mechanism for glycogen phosphorolysis. Glycogen phosphorylase is a dimer of identical 842 residues (97 kDa) per subunit. The active enzyme is phosphorylated at Ser14 and requires pyridoxal phosphate (PLP) for the catalysis. In the glycogen phosphorylase catalyzed reaction, a proton may be initially transferred from pyridoxal phosphate (PLP) to the substrate phosphate which acts as a general acid, protonating the α-(1→4) glycosidic bond that links the terminal glucose to the glycogen chain. Cleavage of that bond leads to the formation of an oxocarbonium ion on the free terminal glucose. Finally the orthophosphate may transfer its proton to PLP and simultaneously mount a nucleophilic attack on the oxocarbonium ion, forming G1P

Hydrolysis on the a-side leaves the phosphate attached to the 5′ position of the adjacent nucleotide, whereas the b- side cleavage yields 3′-phosphate. The enzymes that hydrolyze the internal linkages are known as endonucleases and those that remove the terminal nucleotides are called exonucleases (Table 12.9).

The nuclease (EC3.1.–.–) catalyzed hydrolytic reaction at phosphorus atom proceeds via in-line nucleophilic attack by the base of DNases or the 2′-OH of the ribose attached to the scissile bond in the case of RNases (Heydenreich *et al.*, 1993). The in-line mechanism yields a pentacoordinated oxyphosphorane transition state with bipyramidal geometry having the attacking and departing atoms at apical positions (Figure 12.7).

In addition to common nucleases listed in Table 12.9, cells contain many distinct nucleases, which perform essential functions (Deutscher, 1993). For example, RNases are involved in:

- processing of functional RNA, e.g. removal of introns, separation of RNA from polycistronic transcript and terminal trimming/maturation of RNA
- RNA turnover, e.g. turnover of the –CCA sequence of tRNA, modulation of poly(A) tract of mRNA
- degradation of RNA, e.g. turnover of mRNA, removal of processing products, elimination of damaged/denatured/errored RNA, digestion of rRNA and tRNA during stresses, degradation of extracellular RNA

The restriction nucleolysis of DNA by restriction endonucleases (restriction enzymes), which recognizes specific sequences on the target DNA will be described in the Chapter 13.

TABLE 12.9 Cleavage pattern and specificity of some representative nucleases

Nuclease	DNA/RNA	Cleavage	Specificity
Exonucleases:			
Snake venom phosphodiesterase	Both	a	Start at 3′-end, 5′-NMP products
Spleen phosphodiesterase	Both	b	Start at 5′-end, 5′-NMP products
Endonucleases:			
RNase A (pancreas)	RNA	b	Where 3′-phosphate with pyrimidine, oligo products with Py 3′-PO_4 ends
RNase (*Bacillus subtilis*)	RNA	b	Where 3′-phosphate with purine, oligo products with Pu 3′-PO_4 ends
RNase T_1	RNA	b	Where 3′-phosphate with guanine
RNase T_2	RNA	b	Where 3′-phosphate with adenine
DNase I (pancreas)	DNA	a	Preferably between Py and Pu, nicks dsDNA
DNase II (spleen, thymus)	DNA	a	Oligo products
Nuclease S1	Both	a	Cleave single-stranded but not double-strand nucleic acids

Notes: 1. Abbreviations used are: NMP, any monophosphate nucleotide, Py, pyrimidine base; Pu, purine base.
2. Retroviral RNase H (H for hybrid) which degrades the RNA strand of DNA:RNA hybrids, displays both endo- and 3′-exonuclease activities.

12.8 PROTEOLYSIS AND PROTEIN DEGRADATION

12.8.1 Proteolytic mechanism

Proteins are hydrolytically cleaved at the peptide linkages by proteases (peptidases, EC3.4.–.–) (Beynon and Bond, 2001; Sterchi and Stökerm, 1999). Two classes of peptidases are endopeptidases, which cleave internal bonds (e.g. chymotrypsin, trypsin), and exopeptidases, which hydrolyze the terminal residue of a polypeptide chain (e.g. aminopeptidases, carboxypeptidases). Table 12.10 lists some of the common proteases classified mechanistically according to their characteristics.

The active site of proteases, specifically serine and cysteine proteases, is a cleft consisting of seven subsites, each able to accommodate one amino acid reside of the substrate. Four subsites (S_1 to S_4) are positioned on the acyl side of the scissile peptide bond and three subsites (S_1' to S_3') are on the amino side. The corresponding substrate residues are designated P_1 to P_4 and P_1' to P_3'. The active cleft is located between the two domains, with the catalytically competent Ser or Cys forming the catalytic triad with His and Asp/Asn. All serine proteases have identical tertiary folds consisting of two β-barrels with the catalytic (chymotrypsin- equivalent) Ser195, His59 and Asp102 triad at the interface of the two domains. The common feature includes five enzyme-substrate hydrogen bonds at positions P_1 to P_3 serving to properly juxtapose the scissile peptide bond adjacent to the Ser-His catalytic couple such that the nucleophilic Oγ of Ser195 is accurately positioned for attack. Thus the catalytic mechanism of serine proteases are identical whilst their substrate specificity varies. Many structural determinants controlling specificity reside on surface loops, which surround the extended substrate binding site of serine proteases (Perona and Craik, 1997). The catalytic mechanism of serine proteases has been exemplified for chymotrypsin (Chapter 11). A similar catalytic mechanism for a cysteine protease, papain (KempHuis *et al*., 1984) involves the catalytic residue Cys25 and the other members of the catalytic triad, His159 and Asn175, as shown Figure 12.8.

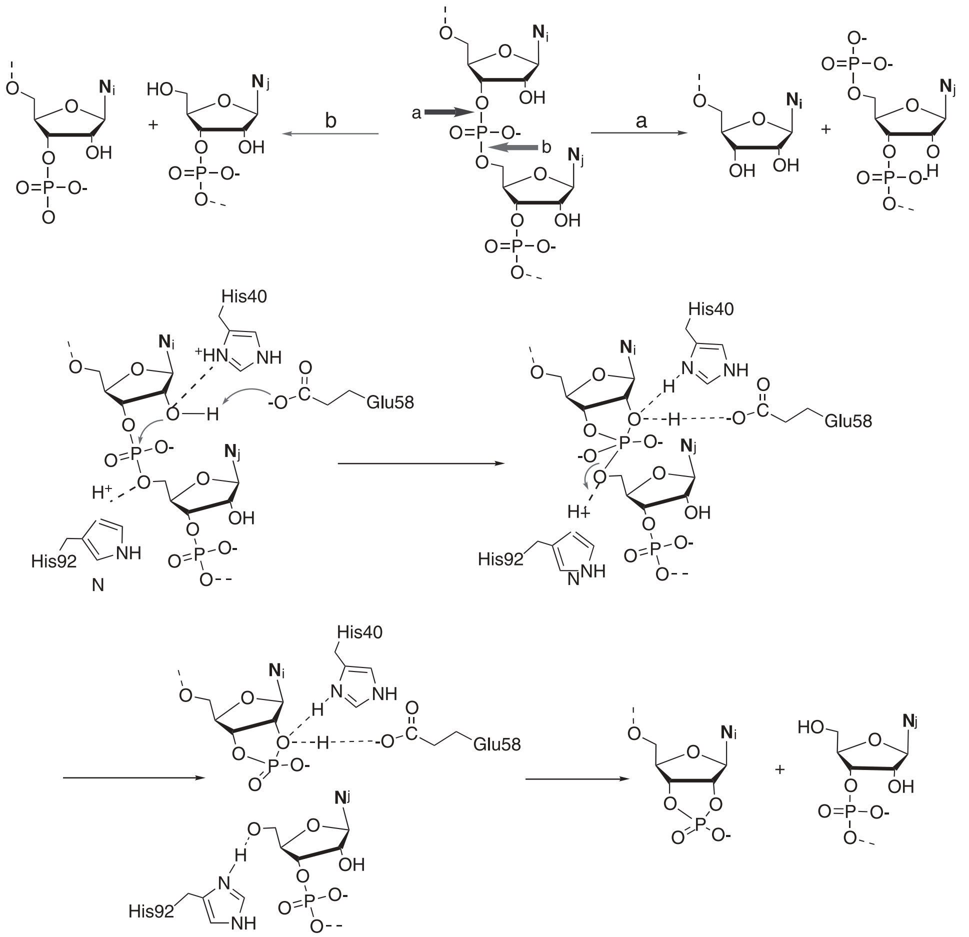


Figure 12.7 Proposed mechanism for RNase T_1 catalysis. RNase T_1 cleaves at N_i = G and the reaction occurs in two steps: (1) a transesterification to yield oligonucleotide with terminal 2′,3′-cyclic GMP and (2) the hydrolysis of 2′,3′-cGMP to 3′-GMP. At optimum pH of catalysis (pH = 4.0 – 7.5), His40 and His92 are protonated and act as acids. Whereas Glu58 is deprotonated and acts as the base, facilitating proton dissociation from the 2′-OH. The catalysis proceeds *via* in-line mechanism with the formation of the pentacoordinated transition-state

Most exopeptidases are metalloproteases (exceptions e.g. D-amino acid aminopeptidase, Salmonella methionine aminopeptidase). Aminopeptidases catalyze the hydrolysis of amino acid residues from the N-terminus of peptide substrates with broad substrate specificity. However, carboxypeptidases hydrolyze C-terminal amino acids with varied substrate specificity. Carboxypeptidase A, which prefers large hydrophobic side chain for the C-terminal residue of peptide substrates, has been extensively investigated (Christianson and Lipcomb, 1989) and its catalytic mechanism is illustrated in Figure 12.9. An analogous mechanism has been proposed for the Zn^{2+} requiring aminopeptidases (Taylor, 1993).

12.8.2 Protein degradation pathway

Most intracellular proteins are in a dynamic state of continued degradation and synthesis with half-lives of proteins ranging from minutes to days. Possible peptide signals that

TABLE 12.10 Examples of proteases classified into mechanistic groups

Family	Active site	Example	Primary function
Serine protease	D-H-S	Chymotrypsin	Digestion
		Trypsin	Digestion
		Thrombin	Blood coagulation
		Plasmin	Lysis of blood clot
		Kallikrein	Control of blood flow
		Elastase	Digestion
		Subtilisin	Digestion
		Acrosin	Sperm penetration
		Cocoonase	Mechanical
		α-Lytic protease	Possibly digestion
Cysteine protease	C	Papain	Digestion
		Cathepsins	Intracellular digestion
		Strept. proteinase	Digestion
Acid protease	D/E	Thermolysin	Digestion
		Pepsin	Digestion
Metallopeptidase	M^{2+}/E,Y	Aminopeptidase	Digestion
		Carboxypeptidase	Digestion

Note: Cathepsins B, H, K, L, and S belong to cysteine superfamily, cathepsin D is an apartyl protease and cathepsin G is a serine protease. Serine protease (http://www.biochem.wustl.edu/~protease) provides sequence, structural and functional information for serine proteases.

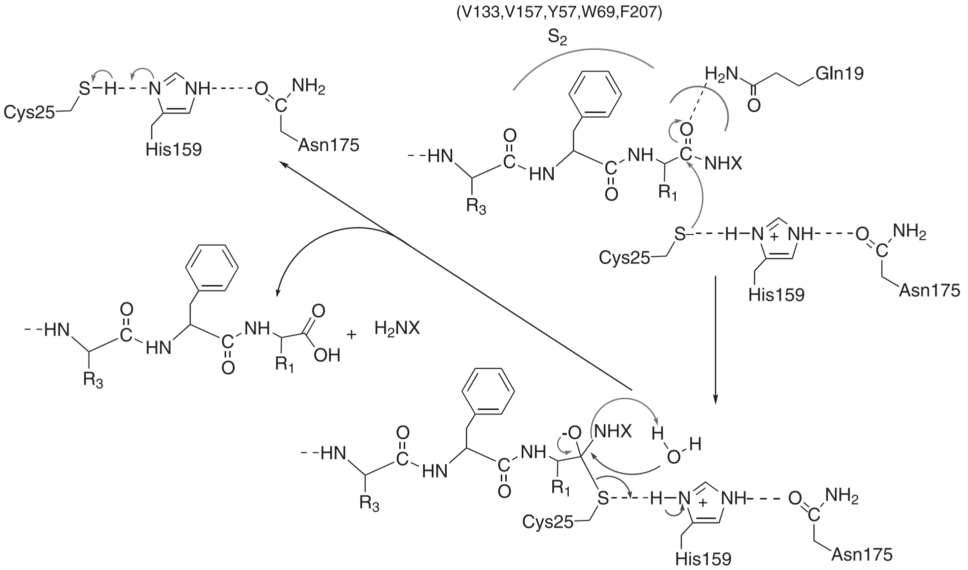


Figure 12.8 Proposed mechanism for papain catalysis. Papain (EC3.4.22.2) consists of a single polypeptide chain of 212 amino acid residues. The protein comprises L-domain and R-domain. The L-domain is mostly α-helices and α-helical turns while the R-domain is based upon a twisted antiparallel β sheets which form a barrel with its interior filled with hydrophobic side chains while two α-helices (residues 117-127 and 138-143) seal opposite ends of the barrel. The primary specificity of papain is for binding a hydrophobic residue in the S_2 subsite (a hydrophobic pocket). This brings the scissile peptide bond to the proximity of the catalytic residue, Cys25 with the polarizable carbonyl group fitting into the oxyanion hole of Gln19. The His159 of the triad abstracts the proton from Cys25 to form a thiolate anion which attacks the polarized carbonyl carbon to form the tetrahedral transition state (there is an uncertainty with respect to whether it is an anion tetrahedral transition state or a protonated tetrahedral transition state). The return of the carbonyl functionality encourages the protonation and cleavage of the amide bond. In the process, the acyl-thioenzyme intermediate is formed and its subsequent hydrolysis releases the product

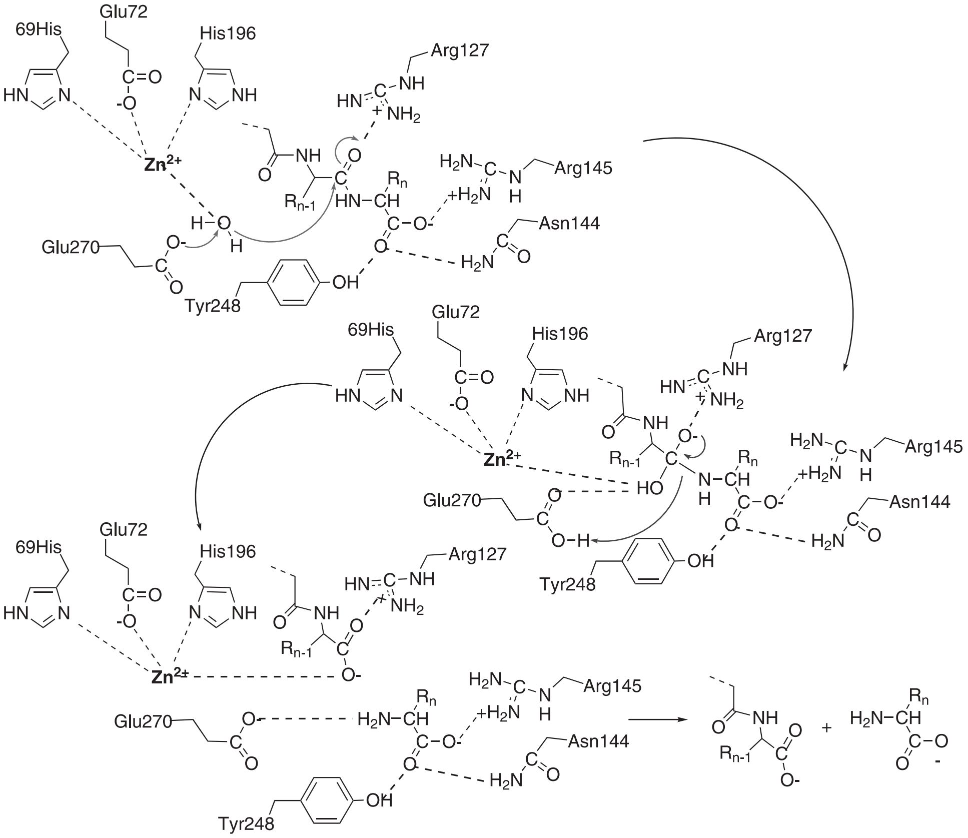


Figure 12.9 Proposed catalytic mechanism for carboxypeptidase A. The C-terminal residue, R_n represent a bulky, hydrophobic side chain. Carboxypeptidase (EC3.3.4.17.–) promotes the polarization of the scissile carbonyl group by hydrogen bonding to Arg127, the activation of water molecule by Zn^{2+} and its deprotonation by Glu270. The zinc-hydroxide ion attack on the carbonyl carbon forms the tetrahedral oxyanion transition state. The formation of products requires protonation of the amino leaving group presumably by Glu270

target proteins for degradation include the PEST region (Rechsteiner and Rogers, 1996), KFERQ motifs and cyclin destruction box (Cdb):

Signal	Sequence	Protein
PEST	HGFPPEVEEQDDGTLPMSCAQES GMDR	Ornithine decarboxylase
	KVEQLSPEEEK	c-Fos
	KDSISPPFAFTPTSSSSSPSPFNSPY K	Yeast cyclin 2
KFERQ	KETAAAKFERQHMDSS	RNase A
Cdb	RTALGDIGN	Cyclin B

PEST sequences are defined as hydrophilic stretches of amino acids greater than or equal to 12 residues that contain at least one each of P, E/D and S/T. The sequences are flanked by K, R or H residues but not interrupted by positively charged residues. Some PEST sequences appear to be constitutive proteolytic signal (e.g. C-terminus of mouse ornithine decarboxylase). However, many PEST sequences are conditional signals and have to be activated (e.g. phytochrome by light, fructose-1,6-*bis*phosphatase by phosphorylation).

Protein degradation is a necessary process for the following reasons:

1. to remove abnormal proteins whose accumulation may be harmful to the cell;
2. to eliminate damaged proteins in the cellular rejuvenation process; and
3. to permit the regulation of cellular metabolism by degradating superfluous enzymes and regulatory proteins.

Cellular proteins are degradated by two major routes, lysosomal and cytosolic ubiquitin (ATP-dependent) pathways. Other likely protein degradations include cytosolic Ca^{2+}-dependent calpains (calpains I and II), cytosolic enzyme that is independent of both ATP and ubiquitin, mitochondria processing, endoplasmic reticulum and plasma membrane proteases (Bond and Butler, 1987).

12.8.2.1 Lysosomal pathway of protein degradation. Lysosomes are membrane encapsulated organelles that contain hosts of hydrolytic enzymes including a variety of proteases known as cathepsins (EC3.4.22.–) (serine proteases for cathepsin R, cysteine proteases for cathepsins B, H, L, M, N, S, T and apartic proteases for cathepsins D, E). Lysosomes recycle intracellular constituents by fusing with autophagic vacuoles (membrane-enclosed portions of cytoplasm) and subsequently breaking down their content. Lysosomal protein degradation in well nourished cells appears to be nonselective, however, a selective pathway may be activated upon prolonged fasting to take over the degradation of proteins containing KFERQ signal sequence (Dice, 1990). KFERQ proteins are bound in the cytosol and delivered to the lysosome by a 73-kDa peptide recognition protein (prp73).

12.8.2.2 Cytosolic ubiquitin (ATP-dependent) degradation. Ubiquitin is a thermostable polypeptide consisting of 76 amino acids. The ubiquitin system (Hershko and Ciechanover, 1998; Pickart, 2001; Glickman and Ciechanover, 2002) is involved in endocytosis and down regulation of receptors and transporters as well as in the degradation of resident or abnormal proteins in endoplasmic reticulum. There are strong indications for roles of the ubiquitin system in developments and apoptosis. In this system, proteins are targeted for degradation by covalent ligation to ubiquitin (Ub). The biochemical steps in the ubiquitin pathway are illustrated in Figure 12.10.

The first step in the ubiquitin pathway is mediated by the Ub-activating enzyme (E_1) to form an E_1-ubiquitin thiol ester between the active site Cys and G76 of Ub via an ubiquitin-adenylate intermediate. E1 is a homodimer with M_r ~210 kDa that is composed of two identical 105 kDa subunits. The yeast enzyme contains two Gly-X-Gly-X-X-Gly motifs that are characteristic of nucleotide-binding domains. Activated Ub is next transferred to an active site Cys residue of Ub-conjugating enzymes or Ubcs (E2s). All E2s share a characteristic UBC domain consisting of ~150 amino acid residues that contains the active site and interacts with Ub, E1 and E3. Specific functions of some E2s may be the result of their association with specific E3s, which in turn bind their specific protein substrates. Stringency of E2–E3 interactions depends on species as well as the identity of the E2 and E3 enzymes. In the third step catalyzed by a Ub-protein ligases (E3s), which have centrally important roles in determining the selectivity of ubiquitin-mediated protein

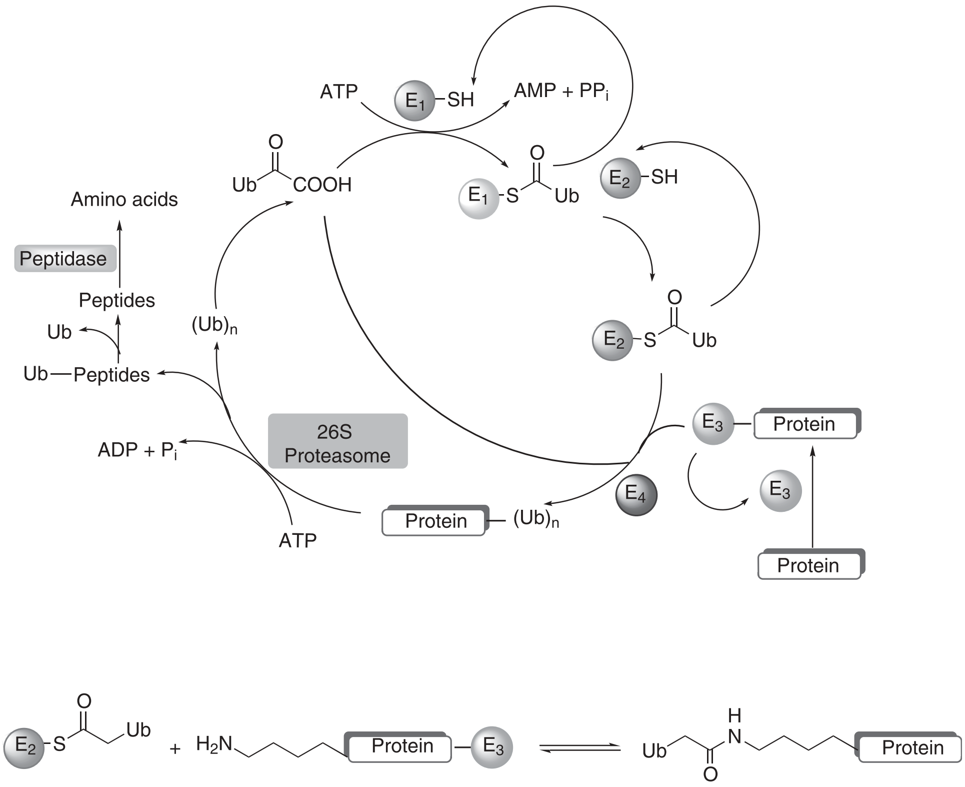
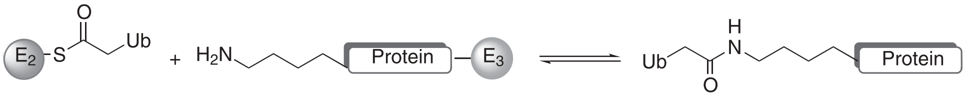

ATP
E1 -SH
AMP + PPi
Ub COOH
O
E1 -S Ub
E2 -SH
Amino acids
Peptidase
Peptides
Ub
Ub—Peptides
(Ub)n
E2 -S Ub
E3
Protein
26S Proteasome
ADP + Pi
ATP
Protein
(Ub)n
E4
E3
Protein
E2 -S Ub
H2N
Protein
E3
Ub
H
N
O
Protein

Figure 12.10 Ubiquitin proteolytic pathway. Proposed sequence of events in conjugation and degradation of proteins *via* ubiquitin system involves activation of ubiquitin (Ub), transfer of activated ubiquitin, conjugation of protein to ubiquitin by ubiquitin-protein ligase, usually polyubiquitination, and degradation of ligated protein by 26S protease complex (proteasome)

degradation, ubiquitin is linked by its C-terminal Gly in an amide isopeptide linkage to an ε-amino group of the substrate protein's Lys residue:


O
E2 -S Ub
+ H2N
Protein
E3
Ub
H
N
O
Protein

After the linkage of Ub to the substrate protein, a polyubiquitin (multiubiquitin) chain is often formed, in which the C-terminus of each ubiquitin unit is linked to a specific Lys residue (most commonly Lys48) of the previous Ub. The multiubiquitin-chain assembly is a processive reaction that usually requires only E1, E2 and E3. However, an efficient multiubiquitination needs an additional conjugation factor termed E4 enzyme (Hoppe, 2005). Ubiquitin-protein ligases are, directly or indirectly, those that bind specific protein substrates, promote the transfer of Ub, and form a thioester intermediate to amide linkages with proteins or polyubiquitin chains.

There are four types of Ub-protein ligases; namely N-end rule E3, *hect*-domain E3, anaphase promoting complex and phosphoprotein-ubiquitin ligase complex. The main N-end rule E3, E3α is the best characterized ubiquitin ligases. It is approximately a 200-kDa protein that binds N-end rule (Varshavsky, 1996) protein substrates that have basic, i.e. Lys, Arg, His (Type I) or bulky-hydrophobic, i.e. Trp, Phe, Tyr, Leu (Type II) N-terminal amino acid residues to separate binding sites. Some protein substrates that do not have N-end rule N-terminal amino acid residues, such as unfolded proteins and some *N*-α-acetylated proteins bind to a putative body site. E3α is responsible for the recognition of some N-end rule protein substrates for Ub ligation and degradation. E3β may be specific for proteins with small and uncharged N-terminal amino acid residues.

A second major family of E3 enzymes is the *hect* domain family. All hect proteins contain a conserved active site Cys residue near the C-terminus. In contrast to the conservation of the C-terminal *hect* domain, the N-terminal region of the different hect proteins are highly variable, probably involved in the recognition of specific protein substrates. A high-molecular-weight complex called cyclosome or anaphase-promoting complex (APC) has a ubiquitin ligase activity specific for cell-cycle regulatory proteins that contain a nine-amino-acid degenerate motif called the destruction box. This complex (~1500 kDa), which is inactive in the interphase but becomes active at the end of mitosis when cyclin B is degraded, exhibits destruction box-specific ubiquitin ligase activity. A different type of multisubunit ubiquitin ligase is involved in the degradation of some other cell-cycle regulators. In these cases, phosphorylation of the substrate converts it to a form susceptible to the action of the ubiquitin ligase complex and therefore designated as phosphoprotein-ubiquitin ligase complexes (PULCs). In addition to the four major types of Ub-protein ligases described above, several other E3s have been found. Because of the variety of mechanisms by which E3 enzymes carry out their two basic functions of protein substrate recognition and Ub transfer, these mechanisms have to be characterized for each types of E3 enzyme. The common property among E3s may be the binding of E2s, but since different E3s bind different E2s with different mechanisms, the major challenge is the identification and elucidation of the mode of action of different E3s that recognize specific signals in cellular proteins called signalosome (Schwechheimer *et al*., 2001).

Protein ligated to polyubiquitin chains is usually degraded by the 26 proteasome complex that requires ATP hydrolysis for its action. The 26S proteasome (2000 kDa) is formed by an ATP-dependent assembly of a 20S proteasome (700 kDa-complex of proteolytic core) with 19S regulatory complex (approximately 20 polypeptides, which determines substrate specificity and provides multiple functions necessary for proteolysis and viability). The 20S proteasomes of prokaryotes and eukaryotes both contain 28 subunits but differ in complexity. Based on sequence similarities, two types of subunits, α and β, are recognized. Prokaryotic proteasomes consist of 14 copies each of the two distinct but related subunits, whereas eukaryotic proteasomes are built of 2 copies each of 7 distinct α-type and 7 distinct β-type subunits. Despite this difference, the other all architecture of these complexes is conserved; i.e. α-type and β-type subunits segregate into 7-member homo- [$\alpha_7\beta_7\beta_7\alpha_7$ (prokarytes)] or hetero-oligomeric [(α1–α7)(β1–β7) (β1–β7)(α1–α7) (eukaryotes)] rings. Two juxtaposed rings of β subunits flanked on the top and bottom by a ring of α subunits form the barrel-shaped complex, which has C_2 symmetry. The barrel-shape particle is 15 nm in height and 11 nm in diameter, consisting of four stacked seven-member rings. The proteolytically-active β-type subunits form the two innermost rings and the α-type subunits form the two outer rings. Together, the four rings enclose three inner compartments, two antechamber and one central proteolytic chamber formed by the β-type subunits. In the yeast 20S proteasome, the entrance to the antechamber is occluded by interdigitating side chain from the N-terminal ends of α-type subunits, suggesting that association with regulatory complexes such as 19S complex is required to render the interior of the complex accessible. The α-type and β-type subunits have similar 3D folds, consisting of two antiparallel five-stranded β-sheet (S1–S10) flanked by two helices (H1, H2) on one side and by three helices (H3–H5) on the other side. In addition, α-type subunits have N-terminal extension (~35 residues) that partly fold into a helix (H0) and fill a cleft in the β-strand sandwich. The cleft is open in the β-type subunits and harbors the active site. The proteasome belongs to the superfamily of N-terminal nucleophile hydrolases (Dodson and Wlodawer, 1998). In the proteasome, the side chain of the amino terminal residue, Thr-1 Oγ of the β-type subunits is the nucleophile for catalytic attack of the carbonyl carbon of a peptide bond. A water molecule and the amino group of Thr-1 are probably required to assist in the extraction of the proton from the side chain hydroxyl to initiate the nucleophilic attack. Target proteins are degraded in a progressive manner by the 20S proteasome without releasing degradation intermediates. Apparently, the central chamber efficiently traps proteins until they are degraded to peptides below a certain length limit, i.e. an average number of 7 to 9 residues.

The assemblage of 26S proteasomes occur by the association of 20S proteasomes with 19S complex (also called μ particle, PA700 or 19S regulator), which caps at either one or both ends of the central 20S complex (Voges *et al.*, 1999). The 26S proteasome has been shown to be the central protease in the ATP- and Ub-dependent degradation of proteins in the eukaryotic cytoplasm and nucleus. It is implicated in the degradation of many rate-limiting enzymes, of abnormal and damaged proteins, of cell-cycle regulators (e.g. cyclins), of oncogens and tumor suppressors, and in the processing of antigens and the activation or degradation of transcription factors. Temporal and spatial control of proteolysis is achieved not only by the selective degradation of ubiquitinated proteins, but also by changes in the cellular localization of the 26S proteasome. By use of specific localization signals, the 26S complex can be deployed to different locations in the cytosol or nucleus, wherever its action is needed.

The action of 26S proteasome generates several types of products including free peptides, and peptides linked to ubiquitin chains, which are converted to free and reusable Ub by the action of ubiquitin-C-terminal hydrolases or isopeptidases. Cytosolic peptidases further degrade short peptides to free amino acids.

12.9 REFERENCES

Ashkenazi, A. and Dixit, W.M. (1998) *Science*, **281**, 1305–8.

Beckerman, M. (2005) *Molecular and Cellular Signaling*, Springer, New York.

Beynon, R. and Bond, J.S. (2001) *Proteolytic Enzymes*, Oxford University Press, Oxford, UK.

Bond, J.S. and Butler, P.E. (1987) *Annual Reviews in Biochemistry*, **56**, 333–64.

Brady, H.J.M. Ed. (2004) *Apoptosis Methods and Protocols*, Humana Press, Totowa, NJ.

Bredt, D.S. and Snyder, S.H. (1994) *Annual Reviews in Biochemistry*, **63**, 175–95.

Burg, D.L., Furlong, M.T., Harrison, M.L. and Geahlen, R.L. (1994) *Journal of Biology Chemistry*, **269**, 28136–42.

Cantrell, D.A. (1996) *Cancer Survey*, **27**, 165–75.

Christianson, D.W. and Lipcomb, W.N. (1989) *Account Chemistry Research*, **22**, 62–9.

Clevenger, E.V. (ed.) (2004) *Signal Transduction*, IOS Press, Burke, VA.

Cohen, P.T. (1997) *Trends in Biochemical Science*, **22**, 245–51.

Coleman, D.E., Noel, J.P., Hamm, H.E. and Sigler, P.B. (1994) *Science*, **265**, 1405–12.

Cory, S. and Adams, J.M. (2002) *Nature Reviews in Cancer*, **2**, 647–56.

Dagley, S. and Nicholson, D.E. (1970) *An Introduction to Metabolic Pathways*, John Wiley & Sons, New York.

De Cesare, D., Fimia, G.M. and Sassone-Corsi, P. (1999) *Trends in Biochemical Science*, **24**, 281–5.

Deutscher, M.P. (1993) *Journal of Biology and Chemistry*, **268**, 13011–4.

Dice, J.F. (1990) *Trends in Biochemical Science*, **15**, 305–9.

Dickson, R.C. and Mendenhall, M.D. (2004) *Signal Transduction Protocols*, Humana Press, Totowa, NJ.

Dodson, G. and Wlodawer, A. (1998) *Trends in Biochemical Science*, **23**, 347–52.

Ellis, L., Hershberger, C.D. and Wackett, L.P. (2000) *Nucleic Acids Research*, **28**, 377–9.

Essen L.-O., Perisic, O., Cheung, R. (1996) *Nature*, **380**, 595–602.

Fletterick, R.J. and Madsen, N.B. (1980) *Annual Reviews in Biochemistry*, **49**, 31–61.

Francis, S.H. and Corbin, J.D. (1994) *Annu. Rev. Physiol.*, **56**, 237–72.

George, S.R. and O'Dowd, B.F. (2005) *G-Protein-Coupled Receptor-Protein Interactions*, Wiley-Liss, Hoboken, NJ.

GIANCOTTI, E.G. (1997) *Current Opinions in Cell Biology*, **9**, 691–700.

GILMAN, A. (1987) *Annual Reviews in Biochemistry*, **56**, 615–49.

GLICKMAN, M.H. and CIEHANOVER, A. (2002) *Physiology Review*, **82**, 373–428.

GOMPERTS, B.D., TATHAM, P.E.R. and KRAMER, I.M. (2002) *Signal Transduction*, Academic Press, San Diego.

GORROD, J.W., OELSCHLÄGER, H. and CALDWELL, J. (eds) (1988) *Metabolism of Xenobiotics*, Taylor & Francis, London.

GYURKO, R., KUHLENCORDT, P., FISHMAN, M.C. and HUANG, P.I. (2000) *Amer. J. Physiol.*, **278**, H971–981.

HAGA, T. and TAKEDA, S. (2005) *G-Protein Coupled receptors: Structure, Function and Ligand Screening* Taylor and Frances, Boca Raton, FL.

HELDIN, C.-H. and PURTON, M. (eds.) (1996) *Signal Transduction*, Chapman and Hall, New York.

HEPHER, J. and GILMAN, A. (1992) *Trends in Biochemical Science*, **17**, 383–7.

HENRISSAT, B. and BAIROCH, A. (1993) *Biochemistry Journal*, **293**, 781–8.

HENRISSAT, B. and DAVIES, G. (1997) *Current Opinions in Structural Biology*, **7**, 637–44.

HERBERT, R.B. (1989) *The Biosynthesis of Secondary Metabolites*, 2nd edn, Chapman and Hall, New York.

HERSHKO, A. and CIECHANOVER, A. (1998) *Annual Reviews in Biochemistry*, **67**, 425–79

HEYDENREICH, A., KOELLNER, G., CHOE, H.W. *et al*. (1993) *European Journal of Biochemistry*, **218**, 1005–12.

HOPPE, T. (2005) *Trends in Biochemical Science*, **30**, 183–7.

HULME, E.C. (ed.) (1990) *Receptor Biochemistry: A Practical Approach*, IRL Oxford University Press, Oxford, UK.

HUNTER, T. (1996) *Biochemistry Society Transactions*, **24**, 307–27.

JACOBSEN, M. and MCCARTHY, N. (eds) (2002) *Apoptosis*, Oxford University Press, Oxford, UK.

JENKIN, L., LO LEGGIO, L., HARRIS, G. and PICKERSGILL, R. (1995) *FEBS Letters*, **362**, 281–85.

JIANG, X. and WANG, X. (2004) *Annual Reviews in Biochemistry*, **73**, 87–106.

KANEHISA, M., GOTO, S., KAWASHIMA, S. and NAKAYA, A. (2002) *Nucleic Acids Research*, **30**, 42–46.

KAZIRO, Y., ITO, H., KOZASA, T. *et al*. (1991) *Annual Reviews in Biochemistry*, **60**, 349–400.

KEMPHUIS, I.G., KALK, K.H., SWARTE, M.B.A. and DRENTH, J. (1984) *Journal of Molecular Biology*, **179**, 233–56.

KIRKEN, R.A., RUI, H., EVANS, G.A. and FARRAR, W.L. (1993) *Journal of Biology and Chemistry*, **268**, 22765–70.

LI, W., DEANIN, G.G., MARGOLIS, B. *et al*. (1992) *Molecular Cell Biology*, **12**, 3176–82.

LINN, S.M., LLOYD, R.S. and ROBERT, R.J. (1993) *Nucleases*, 2nd edn, Cold Spring Harbor Laboratory Press, New York.

LY, H.D. and WITHERS, S.G. (1999) *Annual Reviews in Biochemistry*, **68**, 487–522.

MANN, J. (1987) *Secondary Metabolism*, 2nd edn, Clarendon Press, New York.

MCCARTER, J.D. and WITHER, S.G. (1994) *Current Opinions in Structural Biology*, **4**, 885–92.

MILBURN, M.V. TONG, L., DE VOS, A.M. *et al*. (1990) *Science*, **247**, 939–45.

NACCACHEL, P.H., GILBERT, C., CAON, A.C. *et al*. (1990) *Blood*, **76**, 2098–104.

NEWTON, A.C. (1997) *Current Opinions in Cell Biology*, **9**, 161–7.

NICHOLSON, D.W. and THORNBERRY, N.A. (1997) *Trends in Biochemical Science*, **22**, 299–306.

PERONA, J.J. and CRAIK, C.S. (1997) *Journal of Biology and Chemistry*, **272**, 29987–90.

PICKART, C.M. (2001) *Annual Reviews in Biochemistry*, **70**, 503–33.

PIERCE, K.L. *et al*. (2002) *National Reviews in Cell Molecular Biology*, **3**, 639–50.

PUTNEY, J.W. (ed.) (2005) *Calcium Signaling*, 2nd edn, Taylor and Francis, Boca Raton, FL.

RAVAGNAN, L., ROUMIER, T. and KROEMER, G. (2002) Journ*al of Cellular Physiology*, **192**, 131–7.

RECHSTEINER, M. and ROGERS, S.W. (1996) *Trends in Biochemical Science*, **21**, 267–71.

ROSE, E.M. and WILKIE, T.M. (2000) *Annual Reviews in Biochemistry*, **69**, 759–827.

SAIER, M. Jr. (1987) *Enzymes in Metabolic Pathways*, Harper & Row, New York.

SCHWECHHEIMER, C. SERINO, G., CALLIS, J. *et al*. (2001) *Science*, **292**, 1379–82.

SINGER, W.D., BROWN, H.A. and STERNWEIS, P.C. (1997) *Annual Reviews in Biochemistry*, **66**, 475–509.

SINNOTT, M.L. (1990) *Chemistry Reviews*, **90**, 1171–202.

SONENBERG, N. and GINGRAS, A.C. (1998) *Current Opinions in Cell Biology*, **10**, 268–75.

STERCHI, E.E. and STÖKERM, W. (1999) *Proteolytic Enzymes: Tools and Target*, Springer, New York.

STRADER, C.D., FONG, T.M., TOTA, M.R. and UNDERWOOD, D. (1994) *Annual Reviews in Biochemistry*, **63**, 101–32.

STRASSER, A., O'CONNOR, L. and DIXIT, V.M. (2000) *Annual Reviews in Biochemistry*, **69**, 217–45.

SUN, H. and TONKS, N.K. (1994) *Trends in Biochemical Science*, **19**, 480–5.

SUNAHARA, R.K., DESSAUER, C.W. and GILMAN, A.G. (1996) *Annual Reviews in Pharmalogical Toxicology*, **36**, 461–80.

TANG, W.-J. and HURLEY, J.H. (1998) *Molecular Pharmacology*, **54**, 231–40.

TAYLOR, A. (1993) *Trends in Biochemical Science*, **18**, 167–72.

VARSHAVSKY, A. (1996) *Proceedings of the National Academy of Sciences, USA*, **93**, 12142–9.

VOGES, D., ZWICKL, P. and BAUMEISTER, W. (1999) *Annual Reviews in Biochemistry*, **68**, 1015–68.

YANG, Y. and YU, X. (2003) *FASEB Journal*, **17**, 790–9.

World Wide Webs cited

Apoptosis: http://www.sgul.ac.uk/depts/immunology/~dash/apoptosis
Boehringer Mannheim metabolic chart: http://expasy.houge.ch/cgi-bin/search-biochem-index
Biocatalysis/Biodegradation DB (BBD): http://umbbd.ahc.umn.edu/index.html
Glycosidase classification: http://www.expasy.ch/cgi-bin/list?glycosid.txt.
GPCRDB: http://www.gpcr.org/7tm/
gpDB: http://bioinformatics.biol.biol.uoa.gr/gpDB)
KEGG Pathway Database: http://www.genome.jp/kegg/pathway.html
NRSAS: http://receptors.org/NR/servers/html/
PathDB: http://www.ncgr.org/pathdb
PRED-GPCR: http://bioinformatics.biol.uoa.gr/PRED-GPCR)
Protein kinases: http://www.sdsc.edu/Kinases
Receptor Database: http://impact.nihs.go.jp/RDB.html
Relibase: http://reliase.ebi.ac.uk/
Serine protease: http://www.biochem.wustl.edu/~protease
SEVENS: http://sevens.cbrc.jp/
Signal transduction classification (STCDB): http://bibiserv.techfak.uni-bielefeld.de/stcdb/

CHAPTER 13

BIOSYNTHESIS AND GENETIC TRANSMISSION

13.1 SACCHARIDE BIOSYNTHESIS AND GLYCOBIOLOGY

13.1.1 Biosynthesis of biopolymer: distributive versus processive

Enzyme-catalyzed biosynthesis of a biopolymer may proceed by either distributive polymerization or processive polymerization. In biosynthetic polymerizations, the enzyme that has added a monomeric unit to the growing chain can either dissociate or recombine randomly with other growing termini or it can remain bound to the same chain and increases the chain length by additional units. Enzymes that dissociate between each addition and distribute themselves among all the termini are called distributive. Enzymes that process along the same chain without dissociation are called processive.

13.1.2 Biosynthesis of oligo- and poly-saccharide chains

Carbohydrates are of central significance in the balance between the Earth's 'living' and 'nonliving' carbon, since photosynthesis (Foyer, 1984), which leads primarily to neutral monosaccharides, is largely responsible for reversing the flow from nonliving to living occurring as a result of the normal process of life. At the center of carbohydrate metabolism is D-Glucose (Glc). Once Glc has been formed, different derivatives from the phosphate esters take over the role of providing the driving force for the reactions by which monosaccharides interconvert (Figure 13.1), leading to biosyntheses of oligo- and polysaccharides.

What chemists have found difficult is join two monosaccharides together with a specific glycosidic linkage to the exclusion of all other isomeric products, which nature does routinely. Biosynthesis of homoglycans such as starch and cellulose involve addition of single nucleoside diphosphate residues (e.g. UDP-D-glucose for starch and GDP-D-glucose for cellulose) to an oligo- or polysaccharide chain under the control of the relevant synthases (e.g. glycogen/starch synthase for amylose chain, and cellulose synthase for cellulose), as depicted in Figure 13.2A. Branch points are introduced into the growing polymer by the action of glycosyl-transferases such as 1,4-α-D-glucan (amylo-pectin) branching enzyme (Q-enzyme), which transfers a segment of a (1→4)-α-D-glucan chain to the hydroxyl group at C6 in a similar glucan chain, as shown in Figure 13.2B (Stoddart, 1984).

Biosynthetic processes are endothermic and reductive. Thus glycosidic bond formation in the glycan synthesis requires free energy input ($\Delta G^{\circ\prime} = {\sim}16\,\text{kJ/mol/bond}$). This free energy is usually acquired through the conversion of monosaccharide units to nucleotide phosphoglycoses (NPGs) such as:

Biomacromolecules, by C. Stan Tsai
Copyright © 2007 John Wiley & Sons, Inc.

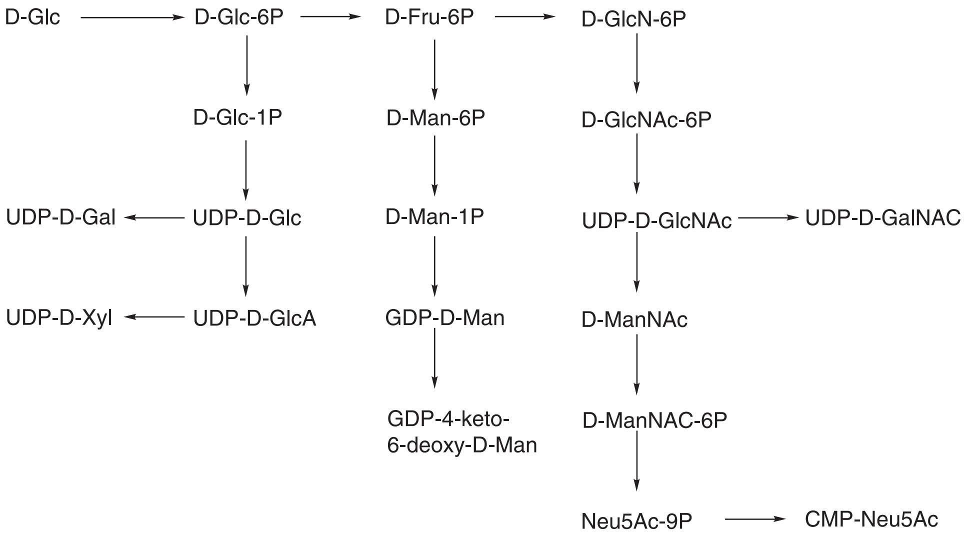


Figure 13.1 Biosynthetic conversion of D-glucose into some of biologically important monosaccharides. The enzymes involved in these monosaccharide interconversions are transferases and isomerases. Nucleotide phosphoglycoses are synthesized as the glycose donors for biosynthesis of glycans. It is noted that 2-amino-2-deoxysugars (glycosamines) arise from D-fructose-6-phosphate and glutamine by the action of D-fructose-6-phosphate:L-glutamine transamidase with subsequent acetylation by acetyl coenzyme A

Nucleotide	Monosaccharide unit
ADP	Glc
CMP	NeuNAc (Sia)
CDP	Glc, Abe (abequose)
GDP	Fuc, Glc, Man
TDP	Glc, Rha (rhamnose)
UDP	Gal, Glc, Xyl, GalNAc, GlcNAc, MurNAc, GluA

Note: *N*-Aceylneuraminic acid = sialic acid, abequose = 3,6-dideoxygalactose, rhamnose = 6-methyl-6-deoxy-L-glucose.

A nucleotide at the anomeric carbon atom is a good leaving group for a saccharide donor and thereby facilitates formation of a glycosidic bond formation to an acceptor via reactions catalyzed by glycosyl transferases. The acceptor commonly has alcohol functionality from another saccharide/saccharide chain or a range of other components of glycoconjugates. The NPG-dependent glycosyl transferases have also been classified by sequence alignments into families (Campbell *et al.*, 1997). The stereochemical outcome of glycosyl transferases, analogous to glycosidases, can be either inversion (e.g. lactose synthetase) or retention (e.g. sucrose synthetase). Either S_N1 with stabilized oxocarbonium ion intermediate or double displacement (two-site insertion) mechanism results in retention, whereas S_N2 (single displacement) mechanism leads to the inversion.

13.1.3 Biosynthesis of glycoproteins

Glycosylation and oligosaccharide processing play an indispensable role in the sorting/targeting/transport of proteins to their proper cellular locations and functions. Cell surface

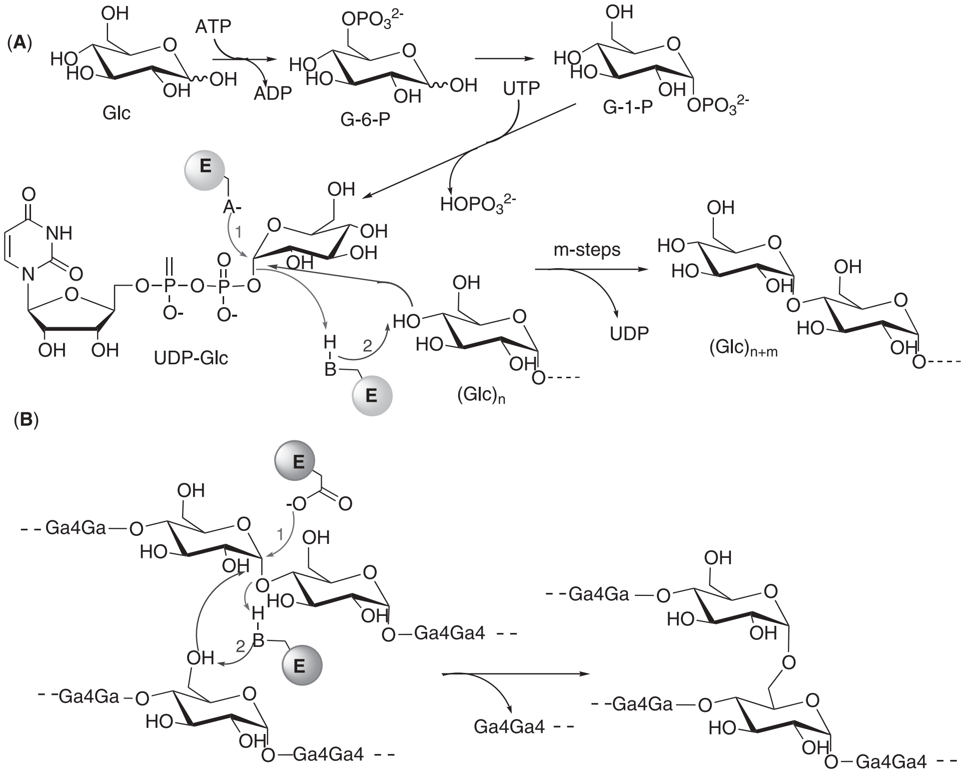


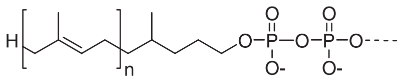
Figure 13.2 Biosynthesis of homoglucan. Biosynthesis of α-homoglucans such as starch/glycogen involves (A) synthesis of linear α-1,4-glucosidic chain (amylose) and (B) formation of branching points, α-1,6-linkages (amylopectin). D-Glc is activated and converted into nucleotide phosphoglucose (e.g. UDP-Glc). Glucan synthase catalyzes the addition of Glc unit from nucleotide phosphoglucose to the nonreducing end of a primer $(Glc)_n$. To form a branching chain, a branching enzyme such as α-1→4/α-1→6-transferase cleaves the α-1→4 glucosidic linkage of the linear chain and transfers a fragment of the chain to the C-6 primary hydroxyl of another chain. The retaining mechanism involves 1, cleavage of the glycosidic bond at C1 with formation of the glycosyl enzyme or the stabilized oxocarbonium ion intermediate and 2, formation of the new glycosidic linkage. The linear code (Ga4 = α-1,4-Glc linkage) is used

carbohydrate groups of glycoproteins mediate a variety of cellular interactions during development, differentiation and oncogenic transformation. Biosynthesis of oligosaccharide chains of glycoproteins is catalyzed by various glycosyltransferases, each specific for creating a single type of linkage (Natsuka and Lowe, 1994). Glycosyltransferases are resident membrane proteins of the endoplasmic reticulum (ER) and the Golgi apparatus. These enzymes have virtually no sequence homology. However, they all have a characteristic topology in the Golgi apparatus. The overall topology consists of short NH_2-terminal cytoplasmic tail, a 16–20 amino acid signal-anchor domain, which spans the membrane, an extended stem region, and a large C-terminal catalytic domain oriented within the lumen of the Golgi cisternae. Biosynthetically, oligosaccharide chains of glycoproteins can be viewed as (a) the core saccharides and their extensions and (b) terminal elaborations (Drickamer and Taylor, 1998). N-linked glycoproteins have a common pentasaccharide trimannosyl core whereas O-linked saccharides are viewed largely as extensions (six types of core structures) and glycosylphosphotidylinositol (GPI)-

membrane anchors have tetrasaccharide core. The synthesis of these oligosaccharide chains will be considered.

13.1.3.1 N-linked glycoproteins. The core saccharide is synthesized as a precursor and transferred *en bloc* to proteins in the ER. The subsequent processing and terminal elaborations take place in the Golgi apparatus. The synthesis occurs in four steps (Kornfeld and Kornfeld, 1985).

Synthesis of a lipid-linked core precursor. N-linked oligosaccharides are initially synthesized as lipid-linked precursors. The lipid component is dolichol (Dol), a long chain polyisoprenol of n = 14–24 isoprene units, which is linked to the oligosaccharide precursor via pyrophosphate bridge.



Dolichol anchors the growing oligosaccharide to the ER membrane. Specific glycosyl transferases catalyze the stepwise addition of monosaccharide units to the growing oligosaccharide chain (Figure 13.3A).

Transfer of the precursor to the carboxamide group of Asn residue on a polypeptide. The dolichol attached core precursor is transferred *en bloc* to the polypeptide chain at Asn residue within the consensus sequence Asn-X-Ser/Thr (X = any neutral amino acid except Pro) in the lumen of the rough ER. The heteroligomeric membrane enzyme, oligosaccharyltransferase (OST) requiring Mn^{2+} for the maximal activity, catalyzes the formation of an N—C bond between the amide nitrogen of Asn and the C1 position of GlcNAc. As such, the transferase reaction can be viewed as a nucleophilic displacement of the Dol-PP from the donor by the carboxamide nitrogen of Asn, which is not inherently reactive nucleophile. The tripeptide sequence of Asn-X-Ser/Thr forms an asparagine-turn conformation. The deprotonation of the amide nitrogen of Asn by the basic residue at the active site of OST may induce tautomerization of the carboxamide to an imidol, which acts as the nucleophile to displace Dol-PP and forms the glycosylcarboxamide linkage.

Removal of some of the precursor's saccharide units. The precursor is trimmed during transit through various luminal compartments and Golgi apparatus. Glc residues and some Man residues are removed by various α-glycosidases, as shown in Figure 13.3B. Following trimming of Glc and some Man residues, the remodeling of the core by the addition of various types of extensions is largely determined by the addition of GlcNAc residues, which dictate various branching patterns. Extensions often include addition of Gal, either as a single residue at the end of a branch or as a longer polylactosamine chain.

Terminal elaborations to the remaining core. The extended structures serve as scaffold for addition of a variety of terminal elaborations. The glycoprotein traverses the Golgi apparatus from the *cis* to the medial to the *trans* Golgi, which contain different sets of glycoprotein processing enzymes. Man residues are then trimmed while Gal, GlcNAc, Man, L-Fuc and Sia (NeuNAc) are added sequentially to form either linear (e.g. attachment of GlcNAc followed by a sulfate group) or branched (e.g. attachment of L-Fuc and Sia) arrangements, giving rise to hybrid and complex types of N-linked oligosaccharides.

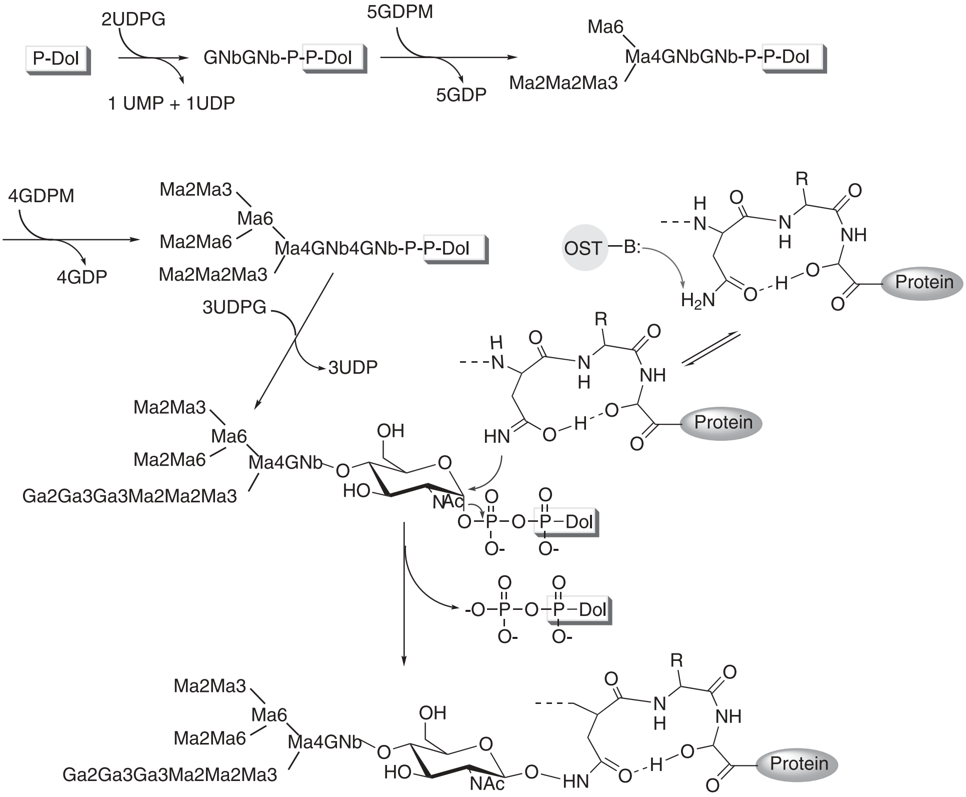

P-Dol
2UDPG
1 UMP + 1UDP
GNbGNb-P-P-Dol
5GDPM
5GDP
Ma6
Ma4GNbGNb-P-P-Dol
Ma2Ma2Ma3
4GDPM
4GDP
Ma2Ma3
Ma6
Ma2Ma6
Ma4GNb4GNb-P-P-Dol
Ma2Ma2Ma3
3UDPG
3UDP
OST —B:
H2N
R
NH
Protein
Ma2Ma3
Ma6
Ma2Ma6
Ma4GNb
Ga2Ga3Ga3Ma2Ma2Ma3
OH
HO
NAc
HN
O-P-O-P-Dol
-O-P-O-P-Dol
Ma2Ma3
Ma6
Ma2Ma6
Ma4GNb
Ga2Ga3Ga3Ma2Ma2Ma3
OH
HO
NAc
HN
Protein

Figure 13.3A Biosynthesis of core oligosaccharides of N-linked glycoproteins: precursors. The core precursor is synthesized as dolichol (Dol)-diphosphooligosaccharide catalyzed by various glycosyl transferases. The addition of two GlcNAc and five Man occur on the cytoplasmic side of the endoplasmic reticulum (ER) membrane. This is followed by a membrane translocation of Dol-PP-(GlcNAc)$_2$(Man)$_5$ to the lumen of ER. Cytosolic GDP-Man and UDP-Glc are converted respectively to Dol-PP-Man and Dol-pp-Glc (not shown) and translocated to the lumen of ER for the subsequent addition to yield Dol attached core precursor, Dol-PP-(GlcNAc)$_2$(Man)$_9$(Glu)$_3$. The core precursor is then transferred to the polypeptide chain at Asn of Asn-X-Ser/Thr. Oligosaccharyltransferase (OST) abstracts the proton to induce tautomerization of the carboxamide to an imidol which acts as a nucleophile to displace Dol-PP. Dolichol diphosphate (pyrophosphate) is translocated to the cytoplasmic surface of the ER membrane and hydrolyzed to dolichol phosphate for re-utilization. Alternately Dol-P can be formed by phosphorylation of Dol by CTP. The linear code for glycose units (see Chapter 6) are used, however the branch connections are explicitly shown for the clarity

13.1.3.2 O-Linked glycoprotein. O-linked oligosaccharides are synthesized in the Golgi apparatus by the stepwise addition of monosaccharide units to the Ser/Thr residue of a completed polypeptide chain, as shown in Figure 13.4. These reactions are catalyzed by specific glycosyl transferases and the Ser/Thr glycosylation sites appear to be influenced by the secondary or tertiary structures of polypeptide chains.

13.1.3.3 Glycosylphosphatidylinositol membrane anchor. Glycosylphosphotidylinositol (GPI) groups function to anchor a wide variety of proteins to exterior surface of the eukaryotic plasma membrane as a transmembrane polypeptide domain. The core GPI is synthesized on the lumenal side of ER (Tartakoff and Singh, 1992), as shown in

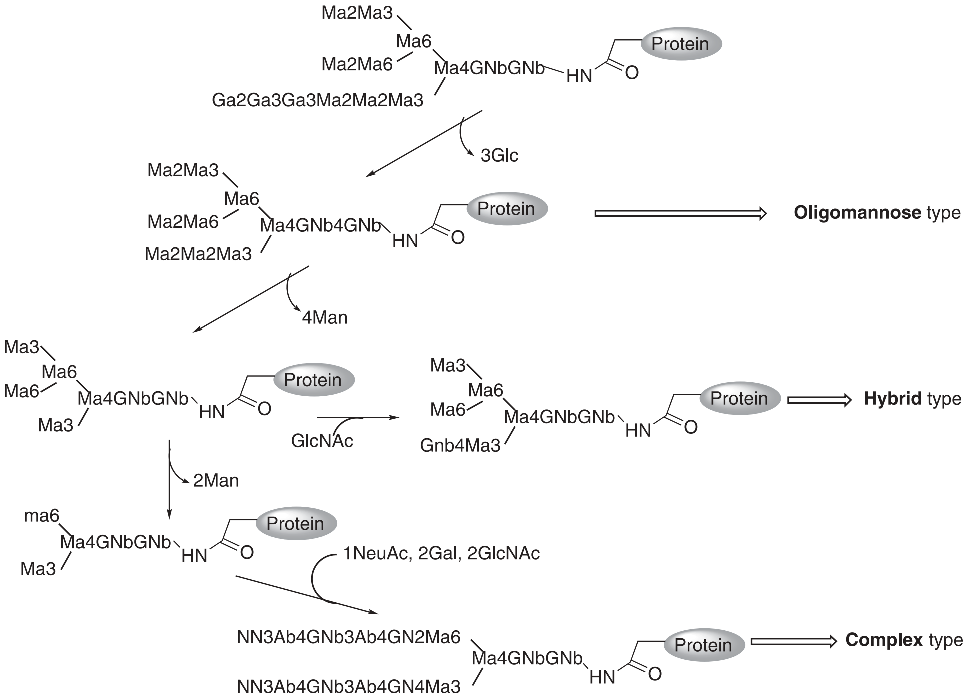

Ma2Ma3
Ma6
Ma2Ma6
Ma4GNbGNb
HN
O
Protein
Ga2Ga3Ga3Ma2Ma2Ma3
3Glc
Ma2Ma3
Ma6
Ma2Ma6
Ma4GNb4GNb
Ma2Ma2Ma3
Protein
Oligomannose type
4Man
Ma3
Ma6
Ma6
Ma4GNbGNb
Ma3
GlcNAc
Gnb4Ma3
Hybrid type
2Man
ma6
Ma4GNbGNb
Ma3
1NeuAc, 2Gal, 2GlcNAc
NN3Ab4GNb3Ab4GN2Ma6
Ma4GNbGNb
NN3Ab4GNb3Ab4GN4Ma3
Complex type

Figure 13.3B Biosynthesis of core oligosaccharides of N-linked glycoproteins: trimming. Glucosidase I (α-1,2) and glucosidase II (α-1,3) hydrolyze outer Glc residues in ER. The product may serves as the core leading to N-linked oligosaccharides of oligomannose (high mannose) type or further demannosylation to serve as the cores for the N-linked oligosaccharides of hybrid as well as complex types. The linear code for glycose units (see Chapter 6) are used, however the branch connections are explicitly shown for the clarity

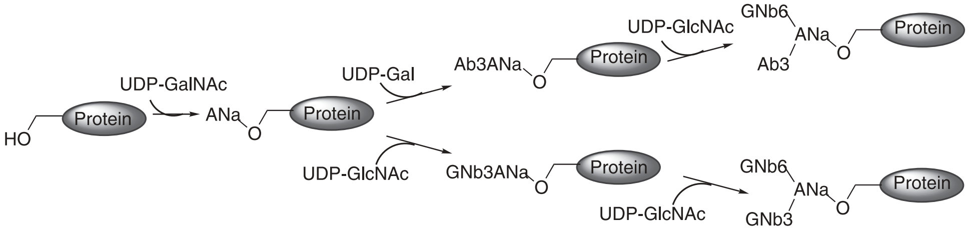

HO
Protein
UDP-GalNAc
ANa
O
UDP-Gal
Ab3ANa
UDP-GlcNAc
GNb6
ANa
Ab3
GNb3ANa
GNb3

Figure 13.4 Possible biosynthetic sequence for O-linked oligosaccharides. The sequential addition of monosaccharide units for the O-linked oligosaccharide synthesis is exemplified for some core structures with an initial N-acetylglucosamine residue. Glycosylation may continue with stepwise addition of Gal, GlcNAc, Fuc and Sia. The linear codes (GNb3/6 for 1,3/6-linkages of β-N-acetylglucosamine and Ab3 for 1,3-linkage of β-galactose) are used

Figure 13.5. The core is modified with various additional monosaccharide units. Target proteins become anchored to the membrane surface when the amino group of the GPI phosphoethanolamine amidates a specific amino acyl group of the protein near its C-terminus, releasing a 20–30 residues of hydrophobic C-terminal peptide. Thus the GPI groups are appended to proteins to the lumenal surface of the rough ER on the exterior surface of the plasma membrane.

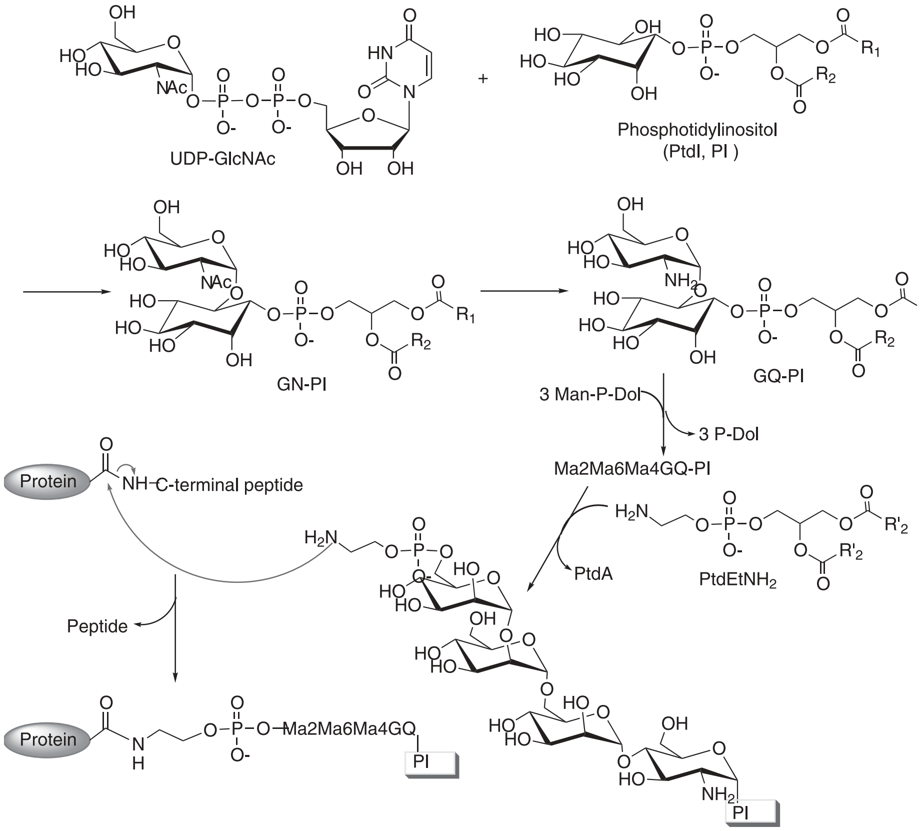


Figure 13.5 Biosynthesis of glycosylphosphotidylinositol protein. The core tetrasaccharide-phosphoinositol structure is synthesized from phosphotidylinositol (PtdI, PI), UDP-GlcNAc, dolichol phosphomannose (Dol-P-Man) and phoshotidylethanolamine ($PtdEtNH_2$). The nucleophilic attack by the amino group of the GPI-phosphoethanolamine forms the GPI-phosphoethanolamine anchored proteins on the surface of the plasma membrane

13.2 GENETIC INFORMATION AND TRANSMISSION

A gene is an inheritable function used to represent a unit of genetic information sufficient to determine an observable trait (Lewin, 1987; Omoto and Lurquin, 2004). In molecular terms, a gene is defined as the DNA sequence necessary to produce a single polypeptide or RNA molecule, while a cistron is the smallest genetic unit that encodes one polypeptide chain.

It is well documented that DNA is the hereditary material (carrier of genetic information), except for some viral RNA. Some of pertinent observations include:

1. DNA is sited exclusively on the chromosome. In prokaryotes, most of the cellular DNA is in the form of a single circular molecule on a single chromosome. However, several species of prokarytes have extrachromosomal small circular DNA molecules (1–30 kbp) known as plasmids, which can reproduce independently. Eukaryotic chromosomes are highly structured complexes of DNA and proteins (histones).
2. The DNA content of a cell is an absolute constant for each species consonant, with the role of DNA as the carrier of genetic information

3. Mutation of genes can be produced by UV light. The action spectrum producing mutations closely corresponds to that of DNA.
4. The purified transforming principle has all the physical and chemical properties of DNA. It is not affected by treatments with trypsin, chymotrypsin or RNase but is completed inactivated by DNase. Thus DNA must be the carrier of genetic information.
5. Most bacterial viruses (bacteriophages) and many animal viruses contain only DNA and protein. In these viruses, transfer of genetic information to progeny virus, known as transduction, has been shown to be due entirely to DNA.

Thus DNA must possess the fundamental properties of the gene, namely:

- It is very stable so that genetic information can be stored within it and transmitted to subsequent generations.
- It is capable of precise copying or replication, so that its information is not lost or altered.
- It is flexible enough to change, in order to accommodate short-term mutation and to adopt long-term evolution.

The DNA duplication and genetic transmission follows:

1. The gene is colinear with the polypeptide it specifies. The flow of genetic information involves biosyntheses of DNA, RNA and proteins known as replication, transcription and translation respectively, as shown schematically in Figure 13.6. It is noted that biosyntheses of DNA, RNA and proteins are template driven, whereas that of glycans is not.

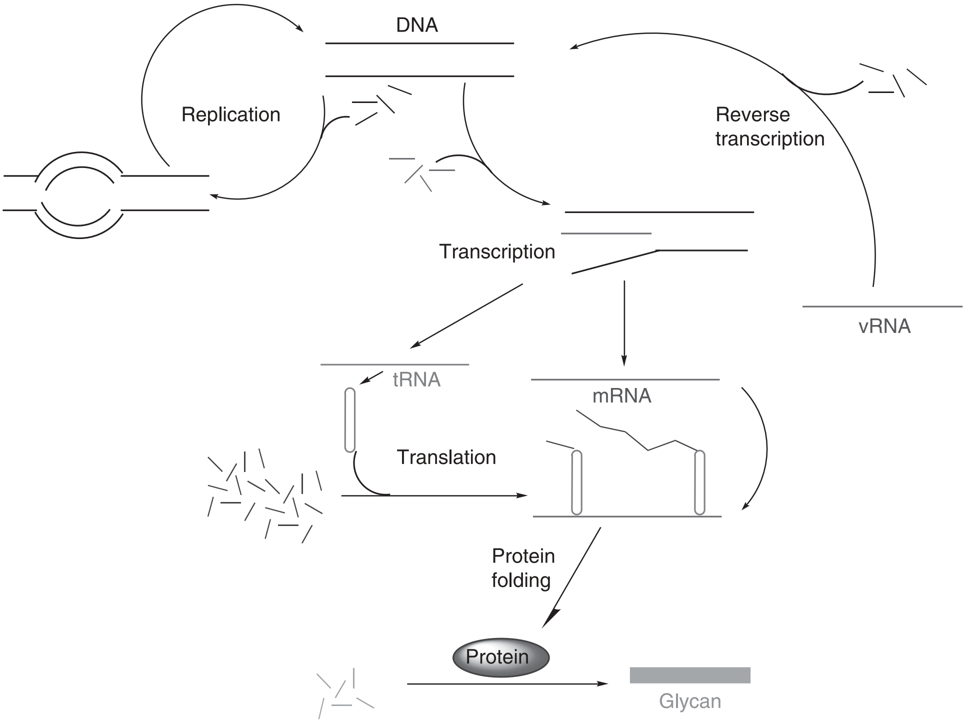


Figure 13.6 Schematic representation of biomacromolecular biosyntheses

2. Nucleic acids and proteins are made of a limited number of different constituent monomers, i.e. four nucleotides for nucleic acids and twenty amino acids for proteins.

3. The monomers are added one at a time in the polymerizations of nucleic acids and proteins. The assembly proceeds step by step in only one chemical direction, i.e. from the 5′ end to the 3′ end in nucleic acids and from amino terminus to the carboxyl terminus in proteins.

4. Each chain has a specific starting point and grows to a fixed terminus; i.e. the start and stop signals are required.

5. Biosyntheses of DNA, RNA and proteins proceed in three phases: initiation, elongation (polymerization) and termination.

6. The primary biosynthetic product is usually modified, i.e. the functional form of a nucleic acid or protein often undergoes post-biosynthetic modification/processing.

Thus DNA replication, RNA transcription and protein translation are more comprehensives in terms that encompass not only nucleotide and peptide chain syntheses, but also their initiation, termination and regulation. The replication copies the parent DNA to form daughter DNA molecules having identical nucleotide sequences. The transcription rewrites parts of the genetic messages in DNA into a form of RNA (mRNA). The translation is the process in which the genetic message coded by mRNA is translated (via carriers, tRNA) into the 20-letter amino acid sequence of protein. In this intricate relationship, DNA directs the synthesis of RNA, then RNA directs the synthesis of protein, but special proteins catalyze the synthesis of both DNA and RNA. So the cyclic flow of information takes place in all cells of living organisms. However, in an infection with retroviruses, viral RNA directs the synthesis of DNA in reverse transcription.

For the nucleic acid that is linear polymer composed of 4 mononucleotide units to specify the linear arrangement of 20 possible amino acids in a polypeptide chain, a group of three nucleotides ($4^3 = 64$) is sufficient and required to encode 20 amino acids. Thus the genetic code consists of triplet (trinucleotide) code words known as codons. The genetic code is read in continuum with stepwise groups of three bases, which is a nonoverlapping, comma-free, degenerate, triple code. The elucidation of the triplet codon for all amino acids permits the translation of nucleotide sequences of DNA into amino acid sequences of proteins (Table 13.1). Genetic codes in various organisms and organelles have been compiled by NCBI at http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi.

All but 3 of the 64 possible codons specify individual amino acids. Because there are 61 codons for 20 amino acids, many amino acids have more than one codon. CUTG at http://www.kazusa.or.jp/codon/ tabulates codon usage from GenBank. The different codons for a given amino acid are synonymous and the code is said to be degenerate, i.e. it contains redundancies. The other three codons are assigned as stop codons (stop signals) that code for the chain termination. They are T(U)AA, T(U)AG and T(U)GA. The three stop codons have inherently different efficiencies. In *E. coli*, UAA is most frequently represented in highly expressed genes. UAG is the less efficient and UGA is naturally leaky and can be recoded for selenocysteine. The initiation codon, AT(U)G is shared with that for methionine. The observation that one kind of organism can accurately translate the genes from entirely different organisms based on the standard genetic code is the basis of genetic engineering. However, the genetic codes of certain mitochondria that contain their own genes and protein synthesizing systems are variants of the standard genetic code (Fox, 1987).

TABLE 13.1 Standard codons

	Middle base				
5′-Terminal base	U(T)	C	A	G	3′-Terminal base
U(T)	Phe	Ser	Tyr	Cys	U(T)
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U(T)
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U(T)
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met(Init)	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U(T)
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

In the translation process, codons in mRNA and anticodons in tRNA interact in an antiparallel manner such that the two 5′ bases of the codon interact with two 3′ bases of the anticodon, according to the AU-pairing and GC-pairing (canonical pairings). However, the 3′ base of the codon may interact with the 5′ base of the anticodon with certain pattern of redundancy known as the wobble hypothesis, i.e.:

3′ Base of codon	5′ Base of anticodon
G	C
U	A
A or G	U
C or U	G
U, C or A	I

13.3 DNA REPLICATION AND REPAIR

13.3.1 DNA replication: Overview

DNA replications in prokaryotes (Marians, 1992) and eukaryotes (Bell and Dutta, 2002; Waga and Stillman, 1998) are similar, though eukaryotic replication is more complex (DePamphilis, 1996). DNA replication has the following general features (Kornberg and Baker, 1992) as shown schematically in Figure 13.7:

1. *Pre-existing DNA as a template*: The DNA chain is produced in cells by the copying of a pre-existing DNA strand according to the Watson–Crick base pairing. The DNA duplex is unwound so that the single strand of DNA molecule may serve as the template. The information in the template is preserved and the copy has a complementary sequence. In the replication of double stranded DNA (dsDNA), both strands of the original duplex are copied.

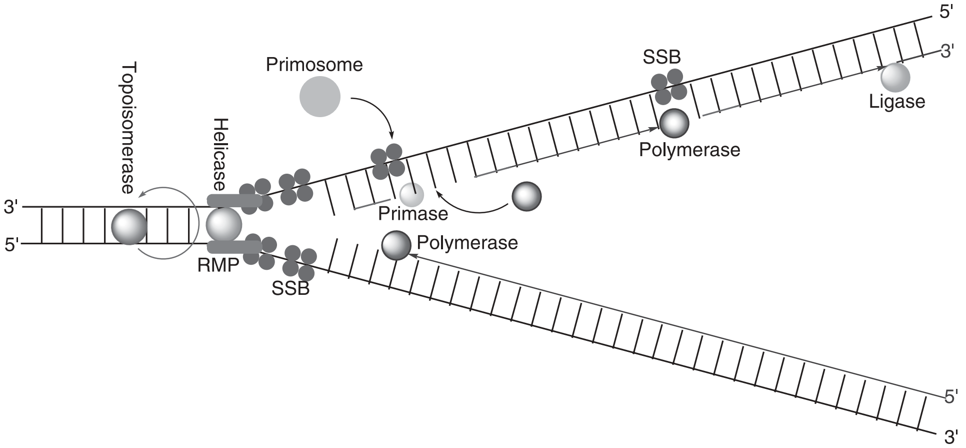


Figure 13.7 Scheme for DNA replication. A prokaryotic replicon is depicted with the associated enzymes and proteins collectively termed DNA replicase system. Both strands of DNA must be replicated. Topoisomerase and helicase (DnaB protein) unwind (ATP requiring) DNA duplex before the advancing replication fork. Single-strand binding protein (SSB) holds the two strands separate. Recombination/replication mediator proteins (RMP) assist the assembly of helicase-SSB complex on ssDNA. In a single-replication fork the synthesis of one strand (the leading strand, continuous arrow), once initiated can proceed in an uninterrupted fashion (e.g. *E. coli* Pol III). The other strand (the lagging strand, broken arrows) must be made (e.g. *E. coli* Pol I) in a retrograde fashion from periodically spaced primers, and the resulting fragments are then stitched together. RNA primers are used and then degraded to result in a complete DNA strand. Similar enzymes/proteins are involved in the eukaryotic DNA replication. The eukaryotic DNA replication fork enzymes/proteins (Waga and Stillman, 1998) consist of Pol α/primase complex, Pol δ/ε, DNA ligase, DNA helicase (primosome assembly), RNase HI (nuclease for removal of RNA primer), RPA (replication protein A for single-stranded DNA binding, stimulation of DNA polymerase and facilitation of helicase loading), RFC (replication factor C i.e. DNA-dependent ATPase for primer-template DNA binding and stimulation of DNA polymerase) and PCNA (proliferating cell nuclear antigen for stimulating DNA polymerases and RFC ATPase). Most replication processes employ a bidirectional pair of replication forks

2. *DNA undergoes semi-conservative replication*: *E coli* cells are grown in a medium containing $^{15}NH_4Cl$ as the sole nitrogen source for several generations to enrich DNA with ^{15}N. These cells are transferred to a fresh medium containing $^{14}NH_4Cl$ and allowed to grow for one doubling time. The DNA, when analyzed with CsCl gradient centrifugation, contains only the hybrid $^{14}N^{15}N$–DNA (the heat-denatured DNA yields two single strands with different densities).

3. *Enzymes*: DNA is replicated by DNA-directed DNA polymerases (Steitz, 1998), which add complementary deoxyribonucleotides to the elongating chain of daughter DNA. DNA ligase is involved in the joining of the nicked DNA fragments. Other enzymes include topoisomerase, helicase and primase, which perform varied functions and are essential to DNA replication. Collectively, these replication enzymes and associated proteins are called the DNA replicase system or the replisome (Benkovic *et al.*, 2001).

4. *Substrates*: All four deoxyribonucleotides) are required for DNA replication. The incoming nucleotide is selected as directed by Watson–Crick base pairing, with the template and triphosphates providing the energy to drive the biosynthesis.

5. *Formation of replication fork*: DNA polymerases use single-stranded DNA (ssDNA) as templates. This involves the unwinding (opening) of the DNA duplex. Circular DNA forms replication eyes or bubbles (θ structures). A branch point in a replication eye at which DNA synthesis occurs is termed a replication fork.

6. *DNA is replicated bidirectionally*: Prokaryotic DNA is a circular duplex. Circular DNA replication involving θ structures is called θ replication, which proceeds in bidirections. Prokaryotic chromosome contains an origin, Ori (a nucleotide sequence of 100–200 base pairs (bp)), where the two DNA replication forks extend in opposite directions.

7. *Requirement of primer*: DNA polymerases cannot initiate a new DNA chain *de novo* (unlike DNA-dependent RNA polymerase, which can find/bind an appropriate initiation site on DNA duplex and begins generating a new RNA strand). An oligonucleotide complementary to the 3-sequence of the template is required as the primer. DNA polymerases only add nucleotide units to the primer chain. RNA primers for Okazaki fragments (see No. 9 below) are synthesized by primase, and are eventually removed and the resulting gaps are filled in with DNA by a 5′ → 3′ exonuclease function of DNA polymerase. Alternative priming involves the specific nicking of one of the strands of a circular DNA providing a free 3′-OH for elongation. In some terminal protein (TP)-containing viruses (e.g. adenovirus, bacteriophage ϕ29), the protein covalently linked to the 5′-end of viral linear DNA duplex provides the OH of Ser, Thr or Tyr as the priming site (Salas, 1991).

8. *Polarity of DNA synthesis*: DNA polymerase adds nucleotide units to 3′-OH of the pre-existing chain. Thus DNA biosynthesis (likewise RNA biosynthesis) proceeds in one chemical direction, from the 5′ to the 3′ end. This directionality has given rise to the convention that polynucleotide sequences are read from left to right in the 5′ → 3′ direction.

9. *Formation of Okazaki fragments*: Pulse-label and pulse-chase experiments (using [^{3}H]thymidine to investigate DNA biosynthesis in *E. coli* culture) demonstrate the formation of DNA fragments consisting of 1000–2000-nucleotides (100–200 nucleotides in eukaryotes) known as Okazaki fragments (Ogawa and Okazaki, 1980). Analysis of Okazaki fragments reveals that their 5′ ends consist of short stretches of oligoribonucleotides (1–60 nucleotides, depending on species).

10. *Continuous versus discontinuous synthesis*: DNA polymerases, the enzymes which synthesize DNA extend the strands only in 5′ → 3′ direction. The DNA duplex is antiparallel. The replication forks would have one strand with the 3′ end (for the 5′ → 3′ strand synthesis) and the other with the 5′ end, yet both strands of DNA replicate simultaneously and semiconservatively. Thus DNA replication is semidiscountinuous. The leading strand that extends the DNA strand from the replication fork in the 5′ → 3′ direction is synthesized continuously. The other strand, the lagging strand, is also synthesized in the 5′ → 3′ direction but discontinuously as Okazaki fragments, which are later covalently joined together by an enzyme, DNA ligase.

11. *Fidelity of DNA replication*: DNA replication proceeds typically with no more than one error in every 10^9 to 10^{10} bp (Echols and Goodman, 1991). Such high accuracy in DNA replication arises from:

a) Cells maintain balanced levels of Dntp.

b) The polymerase reaction itself has extraordinary fidelity via initial binding with the complementary dNTP prior to the covalent formation.

c) The 3′ → 5′ exonuclease function of DNA polymerases detects and eliminates the occasional errors made by the polymerase functions (Goodman *et al.*, 1993).

d) Cooperative function of a battery of enzymes is involved.

12. *Eukaryotic chromosomes with multiple replicons*: Eukaryotic chromosomes have multiple replication origins (replicators) in contrast to the single replication origin of prokaryotic chromosomes. The DNA segments that are each served by the replication origin are called replicons (replication units, generally 3–300 kbp). Thus the prokaryotic chromosome is a single replicon, each eukaryotic chromosome containing many replicons (commonly clusters of adjacent 20–80 replicons), which may be regulated by the tissue-specific and cell cycle-specific factors.

13. *D-Loop replication and rolling-circle replication*: Circular mitochondrial DNA is replicated by a process in which leading strand synthesis precedes lagging strand synthesis. Thus the leading strand displaces the lagging strand template to form a displacement or D loop. In some bacteriophages (e.g. λ and ϕx174 phages), one of the two strands of the circular DNA is first cleaved and new nucleotide units are added to the 3′ end of the broken strand. The growth of the new strand around the circular template continuously displaces the 5′ tail of the broken strand from the rolling template circle. The separated 5′ tail becomes a linear template for synthesis of a new complementary strand.

14. *Termination and telomere*: The binding of the replication termination protein (Tus protein) to the terminus region (τ locus) in prokaryotic chromosome impedes the progression of the replication fork and terminates DNA replication. In eukaryotes, the linear chromosomes terminate with telomeres by the action of telomerase.

13.3.2 DNA replication: Enzymology

A number of enzymes mediate DNA synthesis (Burrell, 1993; Kornberg, 1988). These include helicase and topoisomerase (unwinding and provision of the template), primase (synthesis of primer), DNA dependent DNA polymerase (synthesis of polynucleotide chain), and ligase (joining of Okazaki fragments). These enzymes and proteins form DNA replicase system (Johnson and O'Donnel, 2005), as depicted in Figure 13.6.

13.3.2.1 DNA polymerases. The growth of a DNA chain is catalyzed by DNA polymerase (DNA nucleotidyltransferase, EC2.7.7.7), which requires template, primer, all four dNTP and Mg^{2+}:

$$\text{Primer} + (\text{dATP, dCTP, dGTP, TTP})_{n/4} \xrightarrow{\text{DNA-Template}} \text{DNA} + \text{n (pyrophosphate)}$$

The enzyme promotes the formation of phosphodiesteric bonds by the in-line Mg^{2+}-assisted nucleophilic attack of the 3′-OH of the existing polnucleotide chain at the α-phosphorus atom of the dNTP substrate (Steitz, 1993), as depicted in Figure 13.8. The polymerase (Pol) action is characterized by:

1. *Template*: The nucleotide added is selected according to the sequence of the pre-existing DNA chain serving as the template to which the primer anneals via base-pairing.
2. *Primer*: The nucleotide is added to the 3′-OH of the terminal nucleotide of the RNA primer (~20 nucleotides), which is synthesized by primase.
3. *Deoxyribonucleoside-5′-triphosphate substrates*: All four dNTPs must be present. Triphosphates provide energy necessary for polymerization.
4. *Polarity*: The DNA chain grows in the 5′ → 3′ direction, thus the newly synthesized chain is antiparallel to the template.

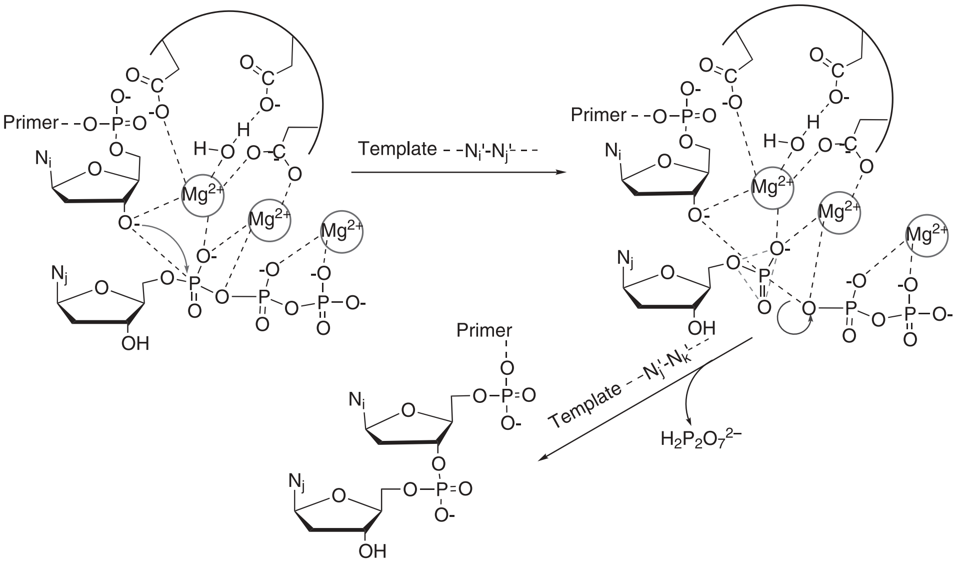


Figure 13.8 Proposed reaction mechanism for DNA polymerase catalysis. The polymerase active site contains three carboxylate residues and probably a lysine. The three carboxylate side chains anchor a pair of divalent metal ions (e.g. Mg^{2+}). In the proposed mechanism, two carboxylates coordinate directly to the two Mg^{2+}, one of which promotes the deprotonation of the 3′-OH of the primer. An in-line attack of the α-phosphorus atom of dNTP forms a bipyramidal pentacoordinated oxyphosphorane transition state with the in-coming and departing atoms at apical positions. The possible involvement of a third Mg^{2+} coordinated to β- and γ-phosphates is also shown

5. *Processivity*: The extent of DNA chain elongation by DNA polymerase per interaction with the DNA primer is referred to as the processivity of Pol. The polymerase is said to be processive if it catalyzes a series of successive polymerization steps.
6. *Fidelity*: The accuracy of incorporating the correct nucleotides per replicon is referred to as the fidelity. Editing is used to correct any misincorporated (mispaired) nucleotides. Some of this editing is carried out by the 3′ → 5′ exonuclease (3′-exonuclease) activity of most Pol.
7. *Nick translation* versus *strand displacement*: Some Pols have a 5′ → 3′ exonuclease (5′-exonuclease) activity that degrades any polynucleotide strand (e.g. primer) in front of a wave of new synthesis. This process is called nick translation. Other Pols that lack this activity will displace one strand of a duplex in order to synthesize a new strand. This process is called strand displacement.
8. *Reversibility*: Reversals of polymerization, such as pyrophosphorolysis of a DNA chain and exchange between pyrophosphate and the β,γ-groups of a dNTP, requires a primer.

DNA polymerases (Pol) have restricted properties so that the information encoded in DNA can be faithfully copied. The properties of prokaryotic polymerases (Kornberg and Baker, 1992) can be summarized as:

1. *Pol I* (*103 kDa*): The three distinct catalytic activities of Pol I, i.e. polymerase (polymerization), 3′-exonuclease (editing) and 5′-exonuclease (primer removal), reside in separate active sites (Joyce and Steitz, 1987). The limited proteolysis with subtilisin or trypsin cleaves the enzyme into two fragments; the smaller fragment (residues

1–323) containing the 5′-exonuclease activity whereas the larger fragment (residues 324–928, known as Klenow fragment) has the polymerase and 3′-exonuclease activities. Pol I has a moderate processivity (~20 nucleotides) and probably involves in the synthesis of the lagging strand and DNA repair. The templates include single strand, nicked duplex and gapped 5′ end DNA.

2. *Pol II* (*90 kDa*): The smallest and least efficient (30 nucleotides/min at 37°C), Pol II possesses polymerase and 3′-exonuclease activities. For polymerization, only gapped 5′ end serves as template.
3. *Pol III* (*165 kDa*): Two activities, polymerase and 3′-endonuclease, are associated with Pol III. The core of this multisubunit enzyme consists of α (130 kDa), ε (27.5 kDa) and θ (10 kDa) subunits. The association of the β subunit (41 kDa) with the γ complex ($\gamma\delta\chi\psi$) confer the core enzyme with high processivity at a rate of nearly 1 kb/sec. Each β_2 sliding clamp tethers a core polymerase to the template accounting for the high processivity of Pol III, which is *E. coli* DNA replicase capable of replicating an entire strand of the *E. coli* genome. Only the gapped 5′ end serves as template for polymerization.

The X-ray crystallographic studies of Klenow fragment of DNA polymerase I (KF-Pol I) in comparison with other polymerases (Joyce and Steitz, 1994) reveal that KF-Pol I consists of two domains. The smaller domain (residues 324–517) contains the 3′-exonuclease site and the larger domain (residues 521–928) contains polymerase active site where the three catalytic carboxylates, Asp705, Asp882 and Glu883 are situated. The polymerase domain can be divided into three subdomains named ‘palm’, ‘finger’ and ‘thumb’. The palm subdomain contains a β-sheet that forms the base of the active cleft (tunnel) and contains the catalytic center, the binding site for the 3′ terminus of the primer and partial dNTP binding site. The finger subdomain is virtually all α-helix to form one wall of the cleft and contributes to the binding/orientation of the template and the other part of dNTP binding site. The flexible thumb subdomain with two long antiparallel and two short α-helices forms the other wall of the cleft and may participate in translocation of the product after dNTP incorporation. The primer-template DNA binds to the catalytic cleft such that the primer strand extends 5′ → 3′ from the exonuclease site toward the polymerase site during polymerization. The nascent 3′ end of the primer strand probably shuttles between the polymerase site (polymerization mode) and the exonuclease site (editing mode) (Steitz, 1993).

Five eukaryotic DNA polymerases are known and their properties are summarized in Table 13.2. The primase subunit of Pol α initiates the synthesis of the leading strand by synthesizing RNA primer followed by the addition of a stretch of oligodeoxyribonucleotide to the primer by the polymerase subunit of Pol α, which has a processivity of only ~100 nucleotides. At this point, replication factor (RFC) carries out polymerase switching by removing Pol α and assembles proliferating cell nuclear antigen (PCNA) in the region of the primer strand terminus. Then Pol δ binds to PCNA and carries out highly processive leading strand synthesis. The synthesis of the lagging strand, which requires frequent priming, is carried out by Pol α. Pol ε is highly processive in the absence of PCNA and has a 3′-exonuclease activity. It is also implicated in the DNA replication and repair. Pol β probably participates in the DNA repair process. Pol γ occurs exclusively in the mitochondrion and presumably replicates the mitochondrial DNA.

13.3.2.2 DNA ligase. DNA ligase seals nicks in double stranded DNA where a 3′-OH and a 5′-phosphate are juxtaposed. The enzymes splice the Okazaki fragments together to form a lagging strand during DNA replication. The ligation, which requires the free

TABLE 13.2 Properties of eukaryotic DNA polymerases

	DNA polymerase (Pol)				
	α	β	γ	δ	ε
Molecular mass (kDa):					
Native	>250	36–38	160–300	170	256
Catalytic core	165–180	36–38	125	125	215
Other subunits	50,60,70		35,47	48	55
Cellular location	Nucleus	Nucleus	Mitoch	Nucleus	Nucleus
3′ Exonuclease activity	−	−	+	+	+
Primase activity	+	−	−	−	−
Processivity	Medium	Low	High	High	High
Fidelity	High	Low	High	High	High
Replication	+	−	+	+	+
Repair	−	+	−		+

Notes: 1. Adapted from Wang (1991)

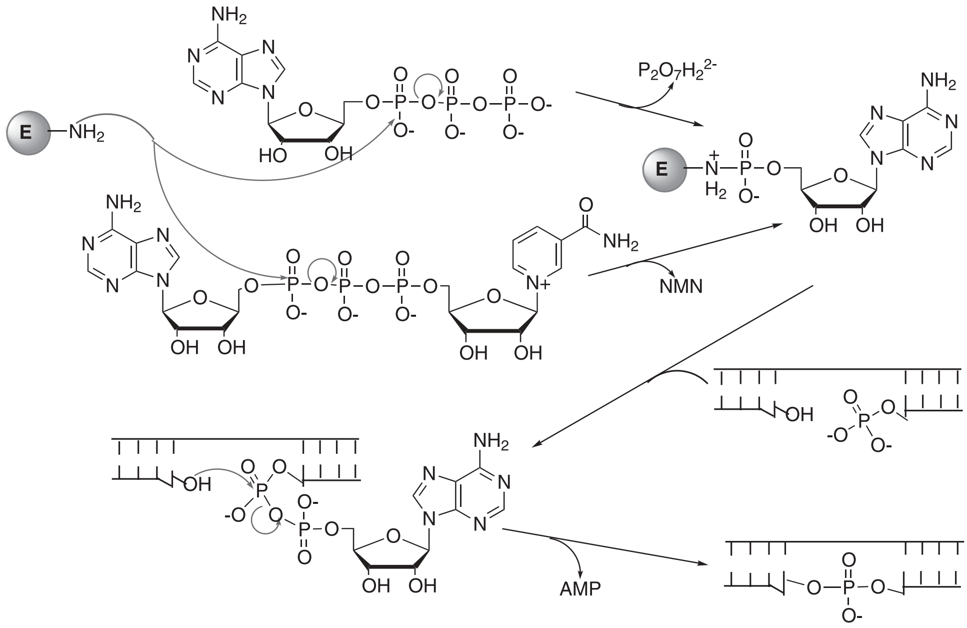
2. Despite the lack of 3′-exonuclease activity, the misincorporation of Pol α is low because of its complexation with primase and the possible presence of a cryptic 3′-exonulease activity. Kinetic sequence of Pol α catalysis is the ordered ter mechanism by interacting, in the order of template (singles strand DNA), primer stem and dNTP.

3. The elongation rate for eukaryotic DNA polymerases is usually 10–15 nucleotides/sec.

energy by the coupled hydrolysis of either ATP (e.g. eukaryotic DNA ligase) or NAD^+ (e.g. *E. coli* DNA ligase), proceeds via formation of adenylyl-enzyme (phosphoramide bond) and adenylated DNA (pyrophosphoryl linkage) intermediates (Figure 13.9). The adenylyl-enzyme intermediates have been isolated for characterization. It is noted that bacteriophage T4 DNA ligase, an ATP-requiring enzyme can link together the DNA duplex known as blunt end ligation at high DNA concentrations. Table 13.3 compares some properties of mammalian DNA ligases (Lindahl and Barnes, 1992).

13.3.2.3 Topoisomerase. Because the two DNA strands are wound about the common helical axis, they are twisted around each other. They cannot unwind without rotation of the structure, once for each helix turn. Thus DNA must spin as it is being replicated. In *E. coli*, for example, the rate of replication is fast enough to make a complete copy of the chromosomal DNA in about 40 minutes. Since the chromosome contains almost 5 million bp of DNA, the required replication forks, each must move at 6×10^4 bases per minutes. Thus each side of the replication fork must unwind 6×10^3 helical turns per minute (it must rotate at 6×10^3 rpm). The negative supercoiling of naturally occurring DNA results in a torsional strain that facilitates unwinding of the duplex helix. The supercoiling of DNA is controlled by a group of enzymes called topoisomerase, which alter the topological state (linking number) of circular DNA (Wang, 1996; Champoux, 2001). Topoisomerases are used to restrict the rotation to regions close to the replication fork. Topoisomerases can cut and reseal dsDNA rapidly without allowing the cut ends to diffuse apart.

Type I topoisomeraes (also known as nick-closing enzymes, monomeric of ~100 kDa) catalyze the relaxation of negative supercoils in DNA by increasing its linking number in increment of one turn. They make a transient single-stranded nick in DNA, which allows one strand to rotate about the other at that point (catenation) and reseal the break, thereby twisting double helical DNA by one turn. *E. coli* type I (IA) topoisomerase interacts both the 5′ and the 3′ side of the site of DNA cleavage and rejoining.



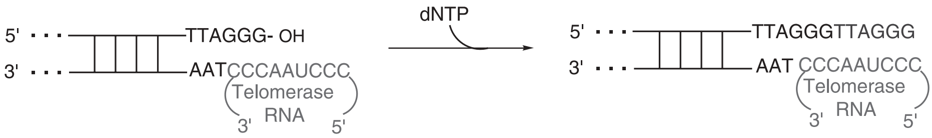


Figure 13.9 Proposed reaction mechanism for DNA ligase. The nick ligation of DNA proceeds in three steps. (1) The adenylyl group of ATP or NAD$^+$ is transferred to the ε-amino group of Lys to form phosphoamide enzyme adduct. In the process, pyrophosphate or nicotinamide monophosphate (NMN) is released. (2) The adenylyl group of this activated enzyme is transferred to the 5′-phosphate terminus of the nick to form an adenylated DNA in which adenylyl monophosphate is linked to the 5′-nucleotide *via* a pyrophosphate bond. (3) The formation of a phosphodiester bond between 3′-OH and 5′-phosphoryl group to seal the nick

TABLE 13.3 Properties of mammalian DNA ligases

	DNA ligase		
	I	II	III
Est. molecular mass (kDa)	125	72	100
K_m for ATP	0.5–1 μM	0.1–0.01 mM	~2 μM
Subcellular localization	Nucleus	Nucleus	Nucleus/cytoplasm?
Induction upon cell proliferation	+	−	−
Ligase activity in calf thymus extract	~85%	5–10%	5–10%
Ligation of oligo(dT)·poly(dA)	+	+	+
oligo(dT)·poly(A)	−	+	+
oligo(A)·poly(dT)	+	−	+

Notes: 1. Adapted from Lindahl and Barnes (1992).

2. DNA ligase I is the key enzyme for joining Okazaki fragments during DNA replication and for completion of DNA excision-repair processes.

Cleavage occurs four nucleotides from the 5′ end and three nucleotides from the 3′ end. The enzyme binds to a negatively supercoiled DNA and does not form a complex with relaxed or positively supercoiled DNA. Eukaryotic type I (IB) topoisomerase binds dsDNA covering a region of ~20 bp. The enzyme is able to remove both negative and positive supercoils.

Type II topoisomerases make a transient double-strand nick (gate) and pass an intact segment of the duplex through this nick. Though it is less obvious, this has an effect of allowing the strands to rotate 720° around each other. When DNA is replicated or transcribed, the double helix must be unwound ahead of and rewound behind the moving polymerases. All type II enzymes interact with a segment of dsDNA centered on the pair of 5′-staggered cleavage sites four bases apart. A 140-bp DNA segment is wrapped around the bacterial type II topoisomerases to form enzyme-DNA complexes. For eukaryotic type II topoisomerases, a binding site spans 20–30 bp of DNA. Parts of the flanking regions within this stretch appear to be exposed to solvent. Prokaryotic type II topoisomerases (~375 kDa also known as DNA gyrases) consist of two pairs of subunits, A and B. DNA gyrase acts by cutting both strands of a duplex, passing the duplex through the break and resealing it. The enzyme catalyzes the stepwise negative supercoiling of DNA with the concomitant ATP hydrolysis and also ties knots in double-stranded circular DNA. By contrast, eukaryotic type II topoisomerases only relax supercoil, they do not generate them nor hydrolyze ATP.

Topoisomerases are recruited for enabling topological motions to occur. In condensed eukaryotic chromatin, loops of 300 Å fiber are attached to a scaffold. Topoisomerases are a major component of the proteins that make up the chromosome scaffold. Although the DNA of mammalian chromosomes is linear, these frequent attachment points make each constrained loop into topologically circular.

13.3.2.4 Primosome: helicase, binding proteins and primase.

Primosome, a complex of enzymes/proteins is required to prime or initiate DNA replication. For the replication of *E. coli* chromosome, DnaA protein (52 kDa) is the initiation factor that recognizes and binds to four 9-bp repeats of the initiation site, oriC. The DnaA protein mediates the separation of the strands of DNA duplex by acting on three AT-rich tandem repeats (5′-GATCTNTNTTNTT) to form the 45-bp open complex. A hexameric DnaB protein (50 kDa with helicase function), delivered by DnaC (assisted by DnaT) binds to the open oriC to form pre-priming complex. ATP or NTP (ribonucleoside-5′-triphosphate) hydrolysis drives the formation of the open and pre-priming complexes. Helicase, with assistance from DNA gyrases, further unwinds the pre-priming complex in both directions. Helicases unwind dsDNA, which is a prerequisite for DNA replication.

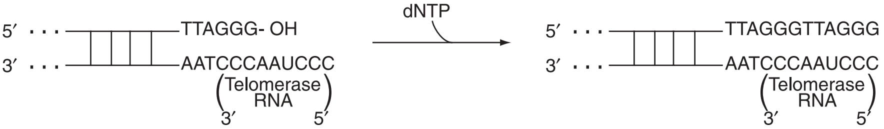
Helicase catalyzed unwinding of dsDNA may proceed by either active mechanism or passive mechanism (Lohman and Bjornson, 1996). In active mechanism, the helicase plays a direct role in destabilizing the DNA duplex. The mechanism requires helicases to possess at least two DNA binding sites. The enzyme would interact directly with the dsDNA at the junction transiently and actively destabilize the duplex via conformational changes triggered by NTP hydrolysis. Alternatively the helicase binds simultaneously to both of the single strands at the ss/dsDNA junction and unwinds by distorting the adjacent duplex region through an NTP-induced conformational change. In the passive mechanism, the helicase facilitates unwinding indirectly by binding to the ssDNA that becomes available through transient disjoining of the duplex caused by thermal fluctuations at the ss/dsDNA junction. Thus ssDNA formed transiently at the junction is trapped by the translocating helicase. The mechanism requires the helicase to interact only with ssDNA and to translocate along ssDNA towards the duplex unidirectionally.

Helicases and topoisomerase are distinct in that topoisomerases alter the linking number of the dsDNA by phosphodiester bond cleavage and reunion, whereas helicases simply disrupt the hydrogen bonds that hold the two strands of DNA duplex together. This is accomplished in a reaction that is coupled with the hydrolysis of a NTP, therefore helicases are also nucleoside-5′-triphosphatases. Some helicases unwind not only DNA duplexes but also DNA-RNA hybrids and RNA duplexes (Matson and Kaiser-Roger, 1990). The unwinding of dsDNA may proceed in a 5′ → 3′ direction (e.g. DnaB, *E. coli* helicases I and III, and mouse ATPase B) or a 3′ → 5′ direction (e.g. *E. coli* helicases II and IV).

Single strand DNA binding protein (SSB) tetramers coat single-stranded regions as they are formed. SSB covers ssDNA, protecting them against nuclease action, preventing their reannealing and configuring them to serve as templates. Unwinding exposes the nucleotide sequence of the strand so that RNA primers can be synthesized by primase. The addition of primase (60 kDa, DnaG protein) completes the formation of the primosome (a complex of ~700 kDa) that form the replication fork and synthesize RNA primer required for DNA synthesis. Once the primosome has primed leading strand synthesis, it remains associated with the lagging strand and periodically primes Okazaki fragment synthesis.

13.3.2.5 Contrahelicase and telomerase. The terminus region (Ter or τ locus) of *E. coli* is a number of short DNA sequences containing a consensus core element, 5′-GTGTGTTGT. Clusters of three or four Ter sequences are organized into two sets inversely orientated with one another capable of blocking the progression of the replication fork when properly aligned with respect to the approaching fork. The binding of the 36 kDa Tus protein (contrahelicase) impedes the progression of the replication fork. Contrahelicase prevents unwinding of DNA duplex and inhibiting the ATP-dependent helicase activity.

Telmers, which are necessary for maintaining chromosomal integrity by protecting against DNA degradation or rearrangement, are short (5–8 bp) repeating tandem G+T rich sequences of $(TTAGGG)_n$ that form protective caps (1–12 kbp) on the ends of eukaryotic chromosomes. This G+T rich strand is longer than the complementary strand and thus the ends of the chromosome have a protruding 3′-end strand of considerable length that folds back on itself to make a stable helical structure. DNA polymerases cannot replicate the very 5′-ends of chromosomes because these enzymes require a template and a primer and synthesize DNA only in 5′ → 3′ direction. Telomeres provide a mechanism by which the ends of the chromosomes can replicated. The RNA primers served in the synthesis of Okazaki fragments of the lagging strand are eventually removed resulting in gaps (primer gaps) in the daughter 5′-terminal strands at each end of the chromosomes. Telomerase is a reverse transcriptase (RNA-dependent DNA polymerase, TERT) of ribonucleoprotein containing RNA component (TERC) of repeating telomeric sequence that serves as a template for the synthesis of telomeres (Blackburn, 1992). For example, nucleotides 46 to 56 (CUAACCCUAAC) of the RNA component (962 ribonuclotides) of human telomerase serves as the template for the DNA polymerase activity of telomerase to add successive TTAGGG repeat units to the 3′-end of a DNA strand to create the G-rich region which in turn create the G-rich region at the 5′-end.



A round of telomeric repeat synthesis in human telomerase involves a six-nucleotide polymerization cycle followed by a translocation. The steric effect either from the telomerase protein or polymerized chain along the template is proposed to define the 5′-end (limit) of the template sequence. A chromosome would be shortened at both ends by the length of an RNA primer with every cycle of DNA replication without the action of telomerase. Essential genes located near the ends of chromosomes would inevitably be deleted by this process. The somatic cells of muticellular organisms lack telomerase activity suggesting that the loss of telomerase function in somatic cells may be the basis for aging in these organisms, perhaps serving as a molecular clock. Furthermore, telomerase activity is often detected in cancer cells in which the repeat telomeric sequences are stable suggesting that telomerase re-activation might be an obligatory step towards carcinogenesis (Hartley and Villeponteau, 1995). TERT of telomerase promotes epidermal stem cell mobilization and proliferation independent of TERC (Calado and Chen, 2006).

Centromeres are DNA regions necessary for precise segregation of chromosomes to daughter cells during cell division. A small repeating sequence $(GGAAT)_n$ has been found to be conserved across a wide range of species.

13.3.3 Reverse transcription

Retroviruses are eukaryotic RNA-viruses, such as certain tumor viruses and human immunodeficiency virus (HIV) containing an RNA-directed DNA polymerase (reverse transcriptase, RT). Viral reverse transcriptase is a multifunctional enzyme possessing three enzymatic activities, which catalyze reactions essential to viral replication (Skalka and Goff, 1993):

1. *RNA-directed DNA polymerase activity*: The retroviral RNA (plus-strand of RNA genome) acts as a template for the synthesis of its complementary DNA (minus-strand) yielding an RNA:DNA hybrid by copying the plus-strand of RNA genome. The DNA synthesis is primed by a captured host cell tRNA, whose 3′-end partially unfolds to bp with a complementary segment of the viral RNA. Reverse transcriptases lack editing function and are error prone. This may contribute to the high mutation rate of retroviruses.

2. *Ribonuclease H activity*: RNases H of RT have both endo- and 3′-exonuclease activities that specifically degrade the RNA chain of DNA:RNA hybrid. The RT-RNase H activity degrades the template genomic RNA and removes tRNA primer. The template RNA is probably degraded as the RNA:DNA hybrid product passes through the RNase H domain, thereby RNase H functions in a polymerase-coupled mode. The RNase H domain from HIV-1 RT has four conserved acidic residues (Asp43, Glu478, Asp498 and Asp549), which compose binding site for two Mn^{2+} and probably to constitute the active site.

3. *DNA-directed DNA polymerase activity*: This activity replicates the minus strand DNA to yield dsDNA after viral RNA genome is degraded. The polymerase domain of HIV-1 RT shares structural similarity with other polymerases, presumably catalyzing the polymerization with the same reaction mechanism (Steitz, 1993; Steitz, 1998).

Reverse transcriptase from HIV-1 is a heterodimer of 66- kDa (p66) and 51-kDa (p51) subunits, in which p51 is derived from the p66 polypeptide by proteolytic cleavage. The p66 subunit consists of two domains: an N-terminal polymerase domain and a C-terminal RNase domain. The polymerase domains of p66 and p51 each contain four subdomains, from N- to C-terminus, finger, palm, thumb and connection. However, the polymerase domains of p66 and p51 are not symmetrically related but rather associated in a sort of head-to-tail arrangement. In p66, the RNase H domain follows the connection (the RNase H domain is excised in p51). The polymerase active site, where three catalytic carboxylate, Asp110, Asp185 and Asp186 are located, is 18–19 nucleotides apart from the RNase H active site where the metal ion is situated. The DNA assumes a conformation that, near the polymerase active site, resembles A-DNA but near the RNase H domain, more closely resemble B-DNA. This is because RT must also bind RNA:DNA hybrids, which have an A-DNA-like conformation. Most of the protein-DNA interactions involve the sugar-phosphate backbone of DNA and residues of p66's palm, thumb and fingers (Joyce and Steitz, 1994). The fingers-subdomain consisting of the mixed α-helix and β-sheet structure is typical for RT, which uses RNA as a template. Different finger subdomain structures characterize the polymerases, which use different templates; all α-helix for polymerases (e.g. KF-Pol I, T7 RNA polymerase) using DNA template versus mixed α-helix and β-structure for reverse transcriptase (e.g. HIV-1 reverse transcriptase) using RNA template. Thus the finger domain may play an important role in determining the template specificity of a polymerase, perhaps by constraining the substrate into the appropriate A-form or B-form conformation at the polymerase active site (Joyce and Steitz, 1994).

13.3.4 Post-replicational modification

13.3.4.1 Methylation. Most cellular DNAs contain small amounts of methylated bases. DNA methylation is essential in mammals, where it is widely distributed within transposable elements, other repeated DNA and coding regions of most functional genes (Freitag and Selker, 2005). The A and C residues may be methylated to form N^6-methyladenine (m^{6}A), N^4-methylcytosine (m^{4}C) and 5-methylcytosine (m^{5}C). These methyl groups project into the major groove of B-DNA where they may interact with proteins. Methyltransferase catalyzes the transfer of methyl group from *S*-adenosylmethionine (SAM) by the reaction mechanism as shown in Figure 13.10.

The methyltransferase from *Haemophilus haemolyticus* methylates its recognition sequence, 5′-GCGC-3′ in dsDNA to yield 5′-G-m^{5}CGC-3′. SAM donates the methyl group to become S-adenosylhomoserine. The methyl transfer via formation of the methylated intermediate is initiated by the nucleophilic thiolate attack of Cys81, assisted by the general acid Glu119. Nearly all base specific interactions are made in the major

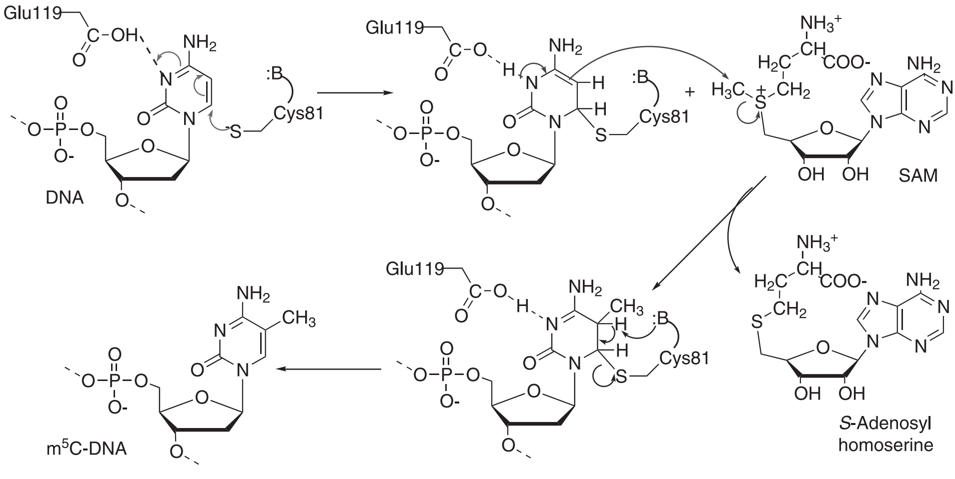


Figure 13.10 Proposed mechanism for 5-methylation of cytosine residue in DNA. Reaction sequences catalyzed by *H. haemolyticus* methyltransferase are depicted. The formation of the methylated DNA intermediate involving thiolate of Cys81 is shown

groove by two so-called recognition loops (residues 233–240 and 250–257) of the enzyme. Methylation of DNA may serve:

1. *Identification of daughter strand in prokaryotes*: Since DNA methylation lags behind DNA synthesis, the daughter strand is under methylated in comparison to the parental strand. The mismatch repair system seeks to act on the undermethylated DNA, which may elude the editing function of Pol during DNA replication.

2. *C_pG island and gene regulation in eukaryotics*: 5-Methylcytosine is the only methylated base in most eukaryotic DNA and occurs largely in the CG dinucleotides of various palindromic sequences. CG is present in the vertebrate genome at only about one fifth of its randomly expected frequency, except for the upstream regions of many genes having normal CG frequency known as CpG islands. Circumstantial evidence implicates that DNA methylation switches off gene expression, particularly in the control regions upstream of the sequences of the transcribed genes.

3. *Role of methylation in mammalian genomic imprinting*: Certain maternally and paternally supplied genes are differentially expressed in mammals, a phenomenon called genomic imprinting and that genomic imprinting is due in part to differential methylation. It appears that parentally supplied genes require a proper level of DNA methylation.

DNA of T-even bacteriophages (e.g. T_2, T_4 and T_6) contains 5-hydroxymethylcytosine (hmC) instead of m^5C and the hmC is generally glucosylated by glucosyl transferase using UDP-Glc as the glucosyl donor.

13.3.4.2 Restriction. The bacterial restriction modification system consists of DNA restriction endonuclease (Pingoud, 2004) and a matched modification enzyme (methylase, i.e. methyltransferase). The restriction endonucleases recognize specific sequences within dsDNA on which the hydrolysis takes place. Three types of restriction enzymes (Table 13.4) have been identified (Yuan, 1981).

Type I and type III RE are bifunctional, carrying both methylation and ATP-requiring restriction activities. The type I enzymes cleaves randomly at the site at least 1 kbp

TABLE 13.4 Characteristics of restriction endonucleases

	Type I	Type II	Type III
Functionality	Bifunctional	Separate functions	Bifunctional
Restriction *versus* methylation	Mutually exclusive	Separate	Simultaneous
Requirement for restriction	ATP, Mg^{2+}, SAM	Mg^{2+}	ATP, Mg^{2+}, SAM
Sequence of host specificity (recognition) sites	Hyphonated, e.g. sB: TGA$(N)_8$TGCT	Twofold symmetry, AluI: AGCT	Disymmetric, N_{5-6}, sPI: AGAGC
Cleavage sites	Random, at least 1 kbp from host specificity site	Identical to host specificity site	24–26 bp toward 3′ of host specificity site
Site of methylation	Host specificity site	Host specificity site	Host specificity site
Enzymatic turnover	No	Yes	Yes
DNA translocation	Yes	No	No
Site of methylation	Recognition site	Recognition site	Recognition site

Notes: 1. Adapted from Yuan (1981).
2. Functionality refers to endonuclease and methylase functionalities.
3. Methylase methylates the amino of A or the 5-position/amino of C using S-adenosylmethionine (SAM) as a methyl donor.

from the recognition sequence, whereas the type III restriction endonucleases cut DNA at specific sites about 24–26 bp from the recognition sequence. By contrast, the type II restriction endonuclesases (commonly referred to as restriction enzymes, REs) with separate methylase, cleave DNA within or near their specific recognition sequence, typically four to six nucleotides in length with a two-fold axis of symmetry known as palindromes. The type II REs that recognize identical sequences are called isoschizomers (e.g. MboI and BamHI). Some type II REs cleave both strands of DNA so as to leave no unpaired bases on both ends, known as blunt ends (e.g. Bal I). Others make staggered cuts on two DNA strands, leaving two to four nucleotides of one strand unpaired at each, the resulting ends referred to as cohesive ends (e.g. Eco RI).

Isoschizomers: Mbo I and Sau IIIA:

↓
5′ . . . G-ATC . . .
3′ . . . C-T-AG . . .
↑

Bal I (blunt end):
↓
5′ . . . T-G-G-C-C-A . . .
3′ . . . A-C-C-G-G-T . . .
↑

Eco RI (cohesive end):
↓
5′ . . . G-A-A-T-T-C . . .
3′ . . . C-T-T-A-A-G . . .
↑

The type II REs that cleave the DNA within the recognition sequences are most widely used for cutting DNA molecules at specific sequences in recombinant DNA and genomic researches. Some of type II REs and their palindromeric recognition/cleavage sequences are given in Table 4.2. The resource site for the restriction endonucleases is accessible at REBASE, http://rebase.neb.com/rebase/rebase.html (Roberts *et al.*, 2005).

13.3.5 DNA repair

Biomacromolecules are susceptible to chemical alternations that arise from environmental and xenometabolic damages or errors during synthesis. For RNA, proteins or glycans, such damages or errors are circumvented by replacement of these molecules through degradation and resynthesis. However, for the integrity of preserving genetic information, any DNA error/damage has to be corrected/repaired. Safeguards include high fidelity replication systems and repair systems. The most common forms of DNA damage are:

a) an incorrect, altered or missing base;
b) deletion or insertion induced bulges;
c) pyrimidine dimer formation;
d) covalent cross-linking of strands; and
e) strand breaks at phosphodiester bonds or deoxyribose rings.

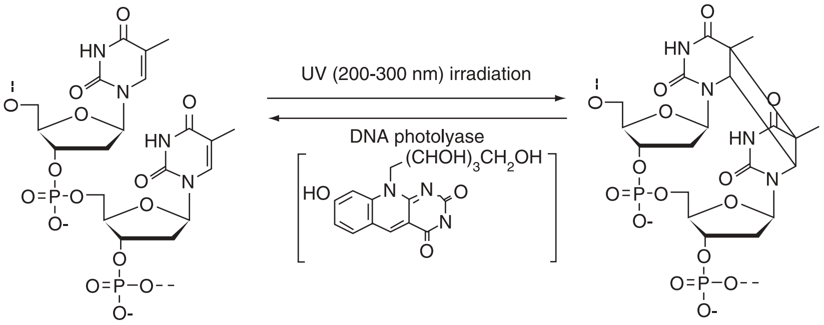
These problems are effectively dealt with by diverse DNA repair systems (Friedberg *et al.*, 1995).

Two fundamental types of cellular responses to DNA damage can be distinguished:

1. Repair of DNA damage including reversal of damage (e.g. enzymatic photoreactivation, alkylated nucleotide repair and ligation of strand breaks) and excision repair; and
2. Tolerance of DNA damage including replicative bypass of template damage with gap formation and translesion DNA synthesis (SOS response). Some of these mechanisms will be considered.

13.3.5.1 Enzymatic photoreactivation of pyrimidine dimmers.

Ultraviolet irradiation promotes the formation of cyclobutyl pyrimidine dimers, which locally distort DNA base pairing structure so that it is unable to form proper replicational or transcriptional template. Pyrimidine dimers may be restored to their monomeric forms by the action of photoreactivating enzymes (DNA photolyases). These enzymes (55–65 kDa) bind to a pyrimidine dimer in DNA, the noncovalently bound chromophore (e.g. N^5,N^{10}methenyltetrahydrofolate or 5-deazaflavin) then absorbs light (300–500 nm) and transfer the excitation energy to a noncovalently bound $FADH^-$, which in turn transfers an electrons to the pyrimidine dimer, thereby splitting it. Finally the resulting pyrimidine anion reduces the FADH· and the unblemished DNA is released.



13.3.5.2 Dealkylation of alkylated nucleotides.

The exposure of DNA to alkylating agents (electrophilic reagents, e.g. methylmethanesulfonate, *N*-methyl-*N*-nitrosourea, *N*-methyl-*N*′-nitro-*N*-nitrosoguanine) yields various alkylated nucleotide derivatives. All N- and O-atoms (except N_1 of pyrimidines and N_9 of purines, which are attached to dexoyribose) of purine and pyrimidine bases are potential alkylating (nucleophilic) sites, though their reactivities may differ. The formation of these derivatives is mutagenic because they frequently cause the incorporation of incorrect nucleotides. For example, the formation of O^6-methylguanine causes the incorporation of thymine instead of cytosine nucleotide. *E. coli* O^6-methylguanine-DNA methyltransferase transfers the offending alkyl group directly to its own Cys residue, thereby rectifying the damage.

13.3.5.3 Nucleotide and base excision repairs.

Pyrimidine dimers and other photoproducts may be removed by a process collectively called DNA excision repair comprising of nucleotide excision repair (NER) and base excision repair (BER). In the NER pathway, the dimer-containing DNA strand at the seventh and fourth phosphodiesteric bonds is cleaved in an ATP-dependent nucleolysis and the excised nucleotides are replaced by the action of DNA polymerase followed by that of DNA ligase.

The BER pathway has evolved to protect cells from the deleterious effects of modified pyrimidine and purine structures. Different DNA glycosylases remove different kinds of base damages by cleaving the glycosidic bonds of the altered nucleotides. Most of these glycosylases are highly specific for a particular form of nonadduct base damage in DNA. Examples include uracil by uracil DNA glycosylase, hypoxanthine (Hx) by Hx DNA glycosylase, 2,5-amino-5-formamidopyrimidine (FaPy) by FaPy DNA glycosylase, urea by urea DNA glycosylase, thymine glycol (Tg) by Tg DNA glycosylase, pyrimidine dimer (PD) by PD DNA glycosylase and methylpurin/pyrimidine (mPu/Py) by mPu/Py DNA glycosylases. Once the base is removed, the apurinic/apyrimidinic (AP)-site is removed by an AP-endonuclease or an AP-lyase, which cleave the DNA strand 5′ or 3′ to the AP-site respectively. The remaining deoxyribose phosphate residue is excised by a phosphodiesterase, and the resulting gap is filled by a DNA polymerase followed by a DNA ligase (Seeberg *et al.*, 1995). For example, uracil residues appearing in DNA as a result of misincorporation of dUTP or deamination of cytosine, can be removed by a uracil DNA glycosylase (glycohydrase) followed by AP-endonuclease and 5′-deoxyribophosphodiesterase. The resulting gap is filled with a cytosine as shown in Figure 13.11.

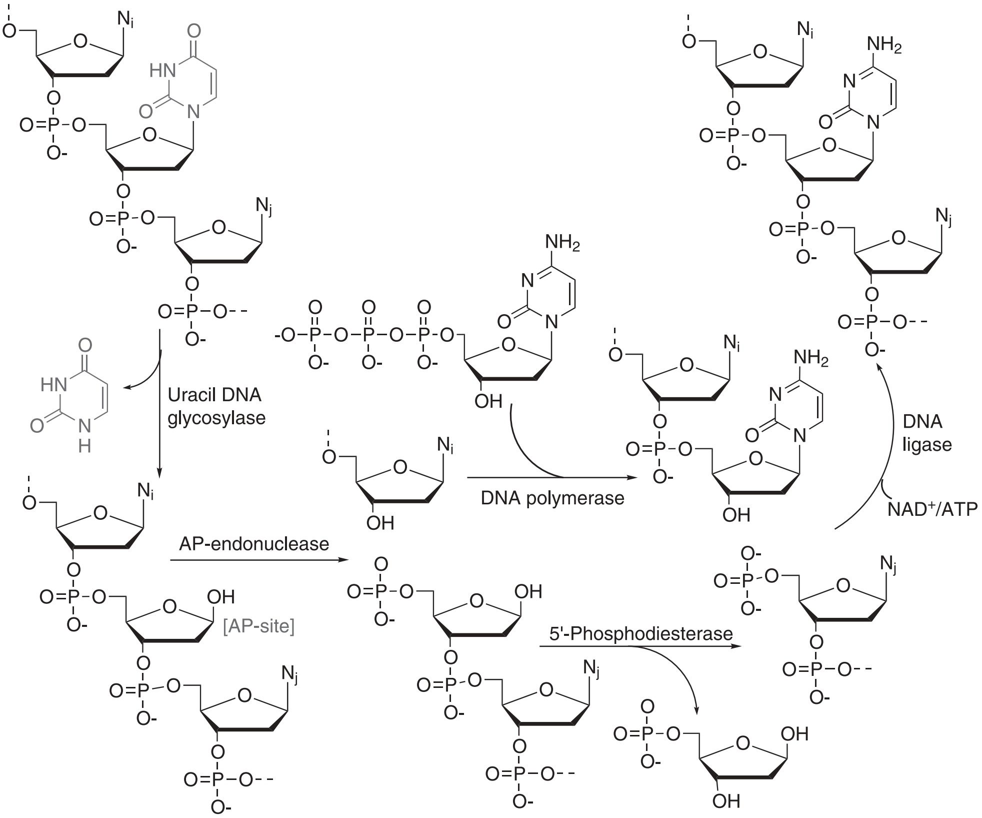


Figure 13.11 Reaction sequence of the base excision repair pathway. The base excision repair (BER) pathway is exemplified for the uracil excision-repair reactions. Only the lesioned (e.g. uracil) part of one strand nucleotide structure of the dsDNA is shown. The excision of uracil by uracil DNA glycosylase without associated AP-lyase generally follows AP-endonuclease and 5′-deoxyribophosphodiesterase which cleave the nucleotide chain. The repair with N-glycosylase is associated with AP-lyase (β-elimination reaction catalyzed by AP-lyase converts the deoxyribose residue to aldehyde form) and 3′-phosphodiesterase. The single nucleotide gap is filled by DNA polymerase (dCTP is required to replace uracil) and DNA ligase

13.3.5.4 SOS translesion DNA synthesis. *E. coli* lesions that block replication such as UV radiation, alkylating or cross-linking agents (SOS mutagenesis), activate the SOS response, a system that converts the lesion to an error prone site to restore replication. This is because the intact SOS response system (SOS mutagenesis system) will replace the bases at a DNA lesion even when there is no information as to which bases were originally present. The SOS signal consists of regions of ssDNA that are generated when the cell attempts to replicate a damaged template or when its normal process of DNA replication is interrupted (Walker, 1995). Thus the SOS response is activated when the replication fork stalls and RecA protein (or recombinase, a 352-residue enzyme that catalyzes the ATP-dependent DNA strand exchange reaction leading to formation of a Holliday junction) binds to exposed ssDNA. The RecA:DNA nucleoprotein filament converts RecA to an activated form (RecA*). LexA protein then interacts with Rec*, changing its conformation to cause autocatalytic proteolytic cleavage at Ala84-Gly85. This cleavage inactivates LexA as a repressor and results in increased expression of many LexA-repressed genes encoding a set of protein mediating error-prone replication and operons in the SOS regulon. These proteins assemble at the lesion to form a mutasome, an error-prone replication apparatus that allows translesion synthesis (Pol III) to occur on damaged DNA template past the lesion.

The rationales for the SOS response system are:

1. The ability to mutate in response to DNA damage confers an evolutionary advantage to the organism.
2. Translesion synthesis occurring under SOS mutagenesis attempts to extract the available base-pairing information from the lesion and, in the case of certain lesions, can usually insert the correct nucleotide.
3. It may be a primary importance for resisting particular classes of environmental agents, which introduce lesions that are not subject to excision repair or other accurate repair systems.

13.4 BIOSYNTHESIS AND TRANSCRIPTION OF RNA

13.4.1 RNA transcription: Prokaryotic system

Cells contain three major classes of RNA, i.e. messenger RNA (mRNA), ribosomal RNA (rRNA), transfer RNA (tRNA), all of which participate in protein synthesis and are synthesized from DNA templates by DNA dependent RNA polymerase in the process termed transcription (Eckstein and Lilley, 1997; Moldave, 1981).

$$(\mathrm{ATP, CTP, GTP, UTP})_{n/4} \xrightarrow{\text{DNA−Template}} \mathrm{RNA} + \mathrm{n}\,(\text{pyrophosphate})$$

Only one RNA polymerase catalyzes the synthesis of all three classes of RNA in prokaryotes. mRNA is transcribed from DNA template carrying genetic codes necessary for protein synthesis. mRNA, which encodes for one protein is called monocitronic, whereas mRNA encoding for more than one protein are referred to as polycitronic. Prokaryotic mRNA is polycitronic and prokaryotic transcription occurs concomitantly with translation.

There are close parallels between prokaryotic and eukaryotic RNA transcriptions. Transcription proceeds in four steps (Figure 13.12), namely:

1. binding of RNA polymerase holoenzyme at promoter site;
2. initiation of polymerization;

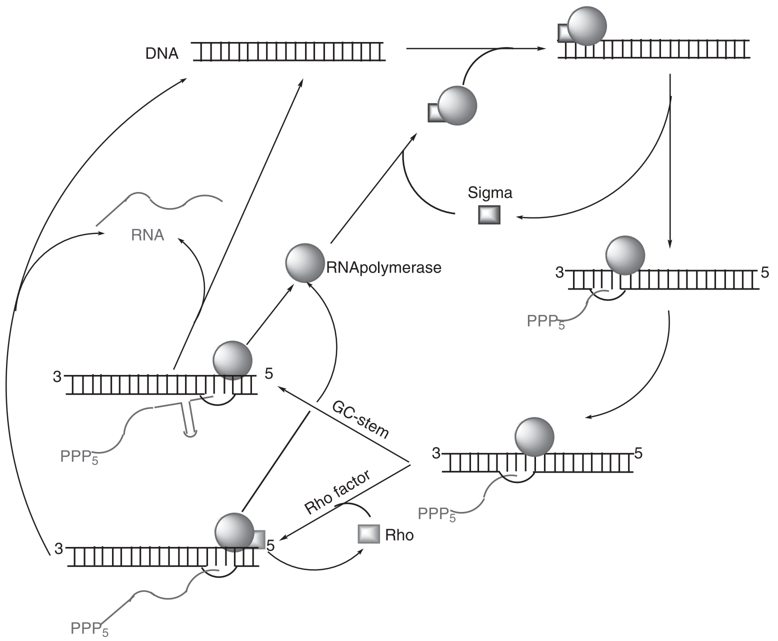


Figure 13.12 Scheme of transcription cycle. Schematic representation of the four transcription steps in prokaryotes is depicted. The elongating RNA chain retains the last 5′-triphosphate. The two alternative termination pathways, GC-rich stem/loop formation and ρ factor participation are shown

3. chain elongation; and
4. chain termination.

The prokaryotic processes pertaining to general features of the RNA transcription will be considered first in this subsection and the eukaryotic processes will be highlighted in the next subsection.

13.4.1.1 Binding of RNA polymerase to DNA template. The transcription begins with the binding of RNA polymerase to form the closed promoter complex through the DNA base sequence known as promoters that are recognized by its σ subunit to form the closed promoter complex ($K_{dissoc} \approx 10^{-6}$–$10^{-9}$ M). Typically prokaryotic promoters consist of a ~40 bp region upstream (− or toward the 5′ direction) of the transcription start site (+1) containing the −10 region with a hexameric TATA box (or Pribnow box) having the consensus sequence of TATAAT and the −35 region having the consensus hexameric sequence, TTGACA. In order for transcription to start, the dsDNA must be opened so that RNA polymerase may gain access to single-stranded template. The RNA polymerase holoenzyme unwinds ~14 bp of DNA to form the stable open promoter complex ($K_{dissoc} \approx 10^{-14}$ M). The RNA polymerase σ subunit is involved in the melting of dsDNA and acts as a sequence specific ssDNA binding protein leaving the template strand available to the catalytic site of the RNA polymerase.

13.4.1.2 Initiation of polymerization. RNA polymerase has two binding sites for NTPs; the initiation site for ATP/GTP and the elongation site for substrate NTPs. Most RNA begins with a purine at the 5′-end. The 3′-OH of the first nucleotide at the initiation site makes a nucleophilic attack on the α-phosphorus atom of the second incoming nucleotide at the elongation site to form a phosphodiesteric bond. Translocation (movement) of RNA polymerase along the template strand prepares the enzyme to add the next nucleotide. Some promoters undergo considerable abortive initiation releasing oligonucleotides of 2–8 residues several times before finally succeeding in producing a long RNA chain. The σ-subunit dissociates from RNA polymerase once oligonucleotide chain of 6–10 residues is formed, signaling the completion of initiation. The core RNA polymerase continues to synthesize the remainder of RNA with about 12 nucleotides of the growing RNA chain remain base-paired to the DNA template at any time.

13.4.1.3 Chain elongation. The elongation (polymerization) of the RNA transcript is catalyzed by the core RNA polymerase at the rate of about 20 – 50 nucleotides/second in the $5' \rightarrow 3'$ direction. As the enzyme moves along the template, the DNA duplex is unwound ahead of the advancing RNA chain and recloses behind the advancing RNA polymerase. It is likely that positive supercoils are created ahead of the transcription bubble and negative supercoils are formed behind it under the mediation of topoisomerases.

13.4.1.4 Chain termination. Two types of transcription termination operate in prokaryotes:

1. A requirement for a termination factor called ρ for transcription termination and RNA release. ρ Factor is an ATP-dependent helicase (hexamer of 46 kDa subunits) that catalyzes the unwinding of RNA:DNA hybrid duplexes or RNA duplex. It binds to C-rich region of RNA, which lacks secondary structure and is unoccupied by translating ribosomes. Upon binding, ρ factor advances in the $5' \rightarrow 3'$ direction, catalyzing the unwinding of the RNA transcript and the DNA template.

2. The termination that is not dependent on ρ, is controlled by the specific sequence in DNA termed termination sites (terminators). The terminators that signal the termination and release of RNA transcript, code for GC-rich inverted repeats punctuated by a nonrepeating segment between them just preceding the 3′ polyU (U_{6-8})-end of the transcript. Therefore a G:C-rich stem/loop structure (hairpin) is formed between the inverted repeats preceding the 3′-polyU end of the transcript. The stem/loop structure causes a pause in the RNA polymerase and resulting dissociation of the RNA transcript form the DNA template.

13.4.2 RNA transcription: Eukaryotic system

Prokaryotic and eukaryotic RNA transcriptions show strong parallels though there are several important differences. A major distinction between prokaryotes and eukaryotes is the move from one prokaryotic enzyme that can faithfully transcribe DNA into RNA to three eukaryotic RNA polymerases. The eukaryotic RNA transcripts are precursors (e.g. pre-mRNA, pre-rRNA and pre-tRNA), which undergo processing to form respective mature RNAs. Furthermore, eukaryotic mRNAs are polyadenylated. A database for mammalian mRNA polyadenylation is available at PolyA_DB (http://polya.umdnj.edu/polyadb). The eukaryotic transcription is tightly regulated and various proteins/factors known as transcription factors (TF) are involved in the eukaryotic transcription. The classification of transcription factors can be found at TRANFAC (http://transfac.gbf.de/TRANFAC/cl/cl.html).

Eukaryotic cells have three distinct RNA polymerases that promotes RNA transcription. RNA polymerase I (RNAPI) is located in the nucleoli and synthesizes the precursor of major rRNAs. RNA polymerase II (RNAPII) is located in the nucleoplasm and synthesizes mRNA precursors (sometimes called heterogeneous nuclear RNA or hnRNA). RNA polymerase III (RNAPIII) is also located in the nucleoplasm and synthesize precursors of tRNA and 5S rRNA. In addition, mitochondrial and chloroplast enzymes are involved in the transcription of mitochondrial and chloroplast genes respectively. Eukaryotic RNA polymerases are large multimeric proteins (molecular weight, 500–700 kDa), which consist of two nonidentical large subunits (>100 kDa) and up to 12 different small subunits (<50 kDa).

The existence of three types of RNAPs acting on three distinct genes implies the existence of at least three categories of promoters. RNAPI only recognizes one species-specific promoter since rRNA genes in a given eukaryotic cell have essentially identical sequences. Mammalian RNAPI requires the presence of core promoter element which span bases −31 to +6. The promoters of RNAPII consist of two separate sequence features; the core element near the transcription site where general transcription factors bind and the distantly located regulatory elements known as enhancers or silencers. The core element often consists of a TATA box (consensus sequence, TATAAA) at −25 region and the transcription start site. The role of the TATA box is to indicate the site of the initiator element (*Inr*) where the transcription is initiated. The regulatory (control) elements are recognized by specific DNA-binding proteins that activate transcription above the basal levels (enhancers) or suppress transcription (silencers). The promoters of RNAPIII can be located entirely within the gene's transcribed regions.

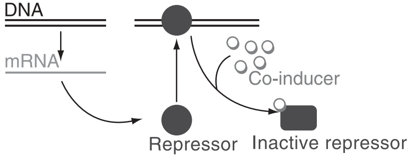
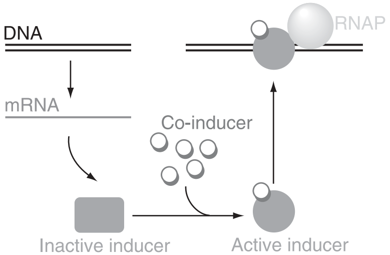
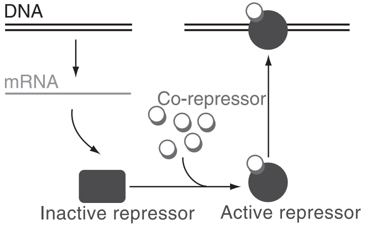
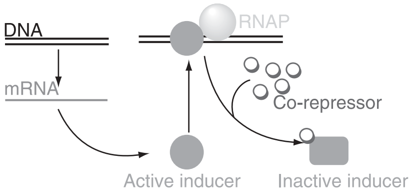
The transcription of mRNA catalyzed by RNAPII involves a universal set of proteins (called basal apparatus), which bind the core promoter to initiate transcription (Table 13.5). The basal apparatus consists of RNAPII and the general transcription factors for RNAPII, as listed in Table 13.5 (Roeder, 1996). TATA-binding protein (TBP) binds to the core promoter through contacts made with the minor groove of the DNA, distorting and bending the DNA so that DNA sequences upstream and downstream of the TATA box come into closer proximity. TFIIB joins, followed by RNAPII in association with TFIIF. Then TFIIE and TFIIH join to establish a competent transcription pre-initiation complex to ensure the formation of open complex forms and the beginning of the transcription.

TABLE 13.5 Human general transcription initiation factors for RNA polymerase II

	Mol. mass (kDa)	Subunits	Function
RNAP II	0–220	12	Catalytic synthesis of RNA transcript, recruitment of TIFF
TIFF	12,19,35	3	Stabilization of TBD binding, stabilization of TAF-DNA interactions, anti-repression functions
TFIIB	35	1	RNAPII-TFIIF recruitment, start-site selection by RNAPII
TFIID:			
TBP	38	1	Core promoter recognition (TATA), TFIIB recruitment
TFA	15–250	12	Core promoter recognition (non-TATA), positive and negative regulatory functions
TFIIE	34,57	2	TFIIH recruitment, modulation of TFII helicase, ATase and kinase activities, promoter melting
TFIIF	30,74	2	Promoter targeting of RNAPII, destabilization of nonspecific RNAPII-DNA interactions
TFIIH	35–89	9	Promoter melting using helicase activity, promoter clearance via CTD phosphorylation

Note: Abbreviations used: CTD, carboxy-terminal domain; RNAPII, RNA polymerase II; TAF, TBP-associated factor; TBP, TATA-binding proteins; TFII, transcription factor for RNAPII.

TABLE 13.6 Control circuits for regulation of prokaryotic transcription

Regulation	Diagram of circuit	Description	Example
Induction, Negative control	 DNA mRNA Co-inducer Repressor Inactive repressor	The substrate (co-inducer) activates synthesis of enzymes by binding to the repressor protein which block the access of RNAP to the promoter.	*lac* Operon: Binding of IPTG to *lac* repressor frees the access of RNAP polymerase to DNA and RNA elongation.
Induction, Positive control	 DNA RNAP mRNA Co-inducer Inactive inducer Active inducer	The inducer protein is activated *via* binding with co-inducer which interacts with the operator to activate synthesis of enzymes (via enhanced transcription)	*lac* Operon: cAMP binding to CAP enhances its DNA-binding ability thereby promoting the formation of RNAP-DNA complex.
Repression, Negative control	 DNA mRNA Co-repressor Inactive repressor Active repressor	The co-repressor-repressor complex binds to DNA to exclude the interaction between RNAP and DNA preventing transcription of the operon.	*trp* Operon: Complex of *trp* repressor-Trp excludes RNAP from the promoter and blocks transcription of *trp* operon.
Repression, Positive control	 DNA RNAP mRNA Co-repressor Active inducer Inactive inducer	The binding of inducer protein promotes RNAP transcription and its removal by the action of co-repressor deters RNAP activity	*ara*BAD Operon (?): Effect on *ara*C in the presence of L-arabinose on AraC protein.

Notes: 1. Abbreviations used: CAP, catabolite gene activator protein (also called cAMP receptor protein, CRP); RNAP, RNA polymerase.
2. Operon Database (DDB) at http://odb.kuict.Kyato-u.ac.jp/ is the data retrieval system for known operons of completed genomes.

13.4.3 Regulation of RNA transcription

Transcription is tightly regulated. In prokaryotes, only about 3% of the genes are undergoing transcription at any given time. In a differentiated eukaryotic cell, it is approximately 0.01%.

13.4.3.1 Prokaryotic regulation. Bacterial genes encoding the enzymes of a particular metabolic pathway are often grouped together in a cluster (operon) with the regulatory sequence (operator) that controls their transcription. The operator is located next to the promoter. The interaction between the operator and a regulatory protein controls transcription of the operon by modulating the accessibility of RNA polymerase to the promoter. Prokaryotic regulation is ultimately responsive to small biomolecules serving as a signal of genetic expression. An increased or decreased synthesis of enzymes in the presence of these signal biomolecules is referred to as induction or repression respectively. The gene products (regulatory proteins), which interact with the operator, are called respectively inducers or repressors and the respective signal biomolecules are called co-inducers or co-repressors. Thus operons can be inducible or repressible, depending on their response to the co-inducers and co-repressors that mediate their expression. Inducible operons are transcribed only in the presence of co-inducers, while repressible operons are expressed only in the absence of their co-repressors. Furthermore, both inducible and repressible operons are subjected to negative and positive control systems. Genes under negative control are transcribed unless they are turned off by the presence of a repressor protein. In contrast, genes under positive control are expressed only if an active inducer protein is present. Table 13.6 summarizes control circuits for prokaryotic regulation (McKnight and Yamamoto, 1992).

Another regulation of RNA transcripts is known as a riboswitch in which RNAs activate or inactivate themselves by changing shape when bound to specific molecules. Regulatory RNAs are located in the 5′ untranslated regions (5′UTRs) of bacterial mRNAs and act as metabolite sensors (Mandal and Breaker, 2004). Ligand binding causes secondary structural changes that lead to regulation of transcription via premature termination or translation by sequestering ribosome-binding sites. For example, bacterial mRNA that encodes thiamine synthesis enzymes in *E. coli* contains a domain that recognizes thiamine pyrophosphate (TPP). In the presence of TPP, this domain changes shape and blocks a nearby sequence, preventing ribosomal interaction and inhibiting translation (Winkler *et al.*, 2002). Riboswitch represents a form of molecular sensing, which is also observed in bacterial ribozyme (Cech, 2004). Ribozyme located in the 5′UTR is activated by the ligand and ribozyme cleavage leads to mRNA degradation and the inhibition of gene expression (Winkler *et al.*, 2004).

13.4.3.2 Eukaryotic regulation. In eukaryotes, the transcriptional regulation is substantially more complicated, such as:

1. The DNA is organized into chromatin, which represses transcription by severely limiting the access of transcriptional regulatory proteins to promoters.
2. Chromatin remodeling involving reversible acetylation of ε-NH_2 of Lys in nucleosomal histones affects gene expression
3. Transcriptional regulation must be coordinated not only for metabolic activities and cell division but also for complex patterns of embryonic development and cell differentiation.

4. The structural genes of eukaryotes are rarely organized in clusters similar to operons. Each eukaryotic gene typically possesses a discrete set of regulatory sequences appropriate to the requirements for its expression.
5. The induction of new proteins in eukaryotic cell takes a much longer time (minutes for prokaryotes but hours or days for eukaryotes) because transcription takes place in the nucleus and the resulting mRNA must be transported to endoplasmic reticulum, where translation occurs (prokaryotic transcription and translation are coupled).
6. Eukaryotic mRNA is relatively more stable and varied. This longer-lived mRNA may have a greater potential for its genetic information to be persistently expressed.

The promoters of eukaryotic genes encoding proteins are defined by modules of short conserved sequences such as the TATA box (TATAAA, ~10 bp at −25 region), CAAT box (GGCCAATCT, ~22 bp at −50 region) and the GC box (GGGCGG, ~22 bp at −80 region). The eukaryotic promoter database is accessible at EPD (http://www.epd-isb-sib.ch). The presence of a CAAT box indicates a strong promoter. One or more copies of the GC box have been found in the region of 'housekeeping genes', which encode proteins commonly present in all cells essential to their normal function. Housekeeping genes are typically transcribed at nearly steady level and sets of various modules are embedded in their upstream region collectively acting as the promoter. The transcription factors that bind to these promoter modules typically behave as positive regulatory proteins.

Eukaryotic genes are characterized by additional regulatory sequences known as enhancers. Like promoters, enhancers represent modules of consensus sequence but differ from promoters in that:

- the location of enhancers relative to the transcription site is varied and act to enhance transcription even if positioned downstream from the gene; and
- enhancer sequences are bidirectional in that they function in either orientation.

Enhancers are promiscuous because they stimulate transcription from any promoter that happens to be in their vicinity. However, enhancer function is dependent on recognition by a specific transcription factor, which by binding to an enhancer element, stimulates RNAPII binding at a nearby promoter via a looping mechanism (Figure 13.13). Eukaryotic transcription must respond to a variety of regulatory signals and multiple proteins are essential for appropriate regulation of gene expression. DNA looping provides dimensions for additional proteins to convene at the initiation site and to exert their influence on forming/activating RNAPII. DNA looping greatly expands the repertoire of transcriptional regulation.

Many eukaryotic genes are subject to a multiplicity of regulatory influences. Promoter modules in genes responsive to such regulations are termed response elements. These elements are found in the promoter region of genes whose transcription is activated in response to physiological challenges. Examples include heat shock element (HSE) to elevated temperatures, glucocorticoid response element (GRE) to steroid hormone concentrations and metal response element (MRE) to toxic heavy metal ions *via* respective transcription factors.

13.4.3.3 RNA interference. Another regulatory mechanism is RNA interference (RNAi), which interferes with the expression of mRNA rather than its biosynthesis (Downward, 2004), as shown in Figure 13.14. Two classes of short RNA molecules, small interfering RNA (siRNA) and microRNA (miRNA) have been identified as sequence

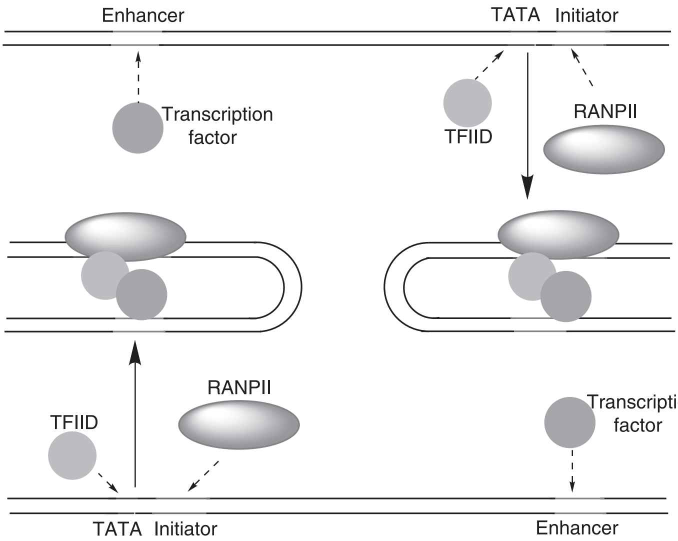

Enhancer
TATA Initiator
Transcription factor
TFIID
RANPII
RANPII
TFIID
Transcription factor
TATA Initiator
Enhancer

Figure 13.13 Enhancer promoter interaction *via* protein mediated DNA loop. Enhancers are sequence elements located at varying positions, either upstream (top) or downstream (bottom). The specific transcription factor (TF) interacts with the enhancer to form DNA looping which delivers the enhancer bound TF to RNAPII positioned at the promoter. The DNA looping enhances transcription complex formation and activates transcription

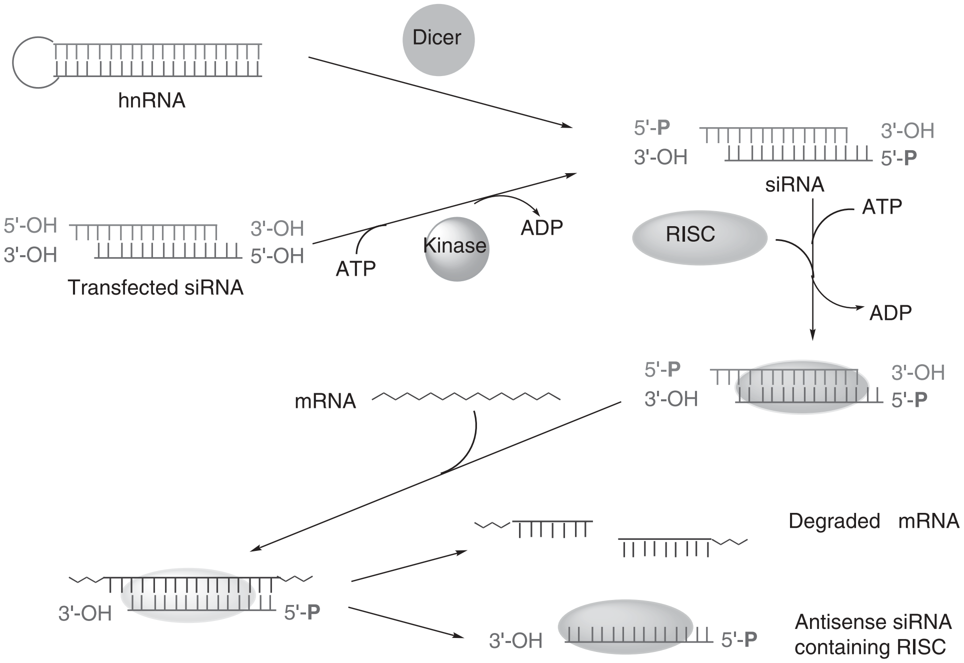

Dicer
hnRNA
5'-P
3'-OH
3'-OH
5'-P
siRNA
5'-OH
3'-OH
3'-OH
5'-OH
Transfected siRNA
ATP
Kinase
ADP
RISC
ATP
ADP
5'-P
3'-OH
3'-OH
5'-P
mRNA
Degraded mRNA
3'-OH
5'-P
3'-OH
5'-P
Antisense siRNA containing RISC

Figure 13.14 Schematic representation of RNA Interference. Small interfering RNA (siRNA) is generated by Dicer (RNase) cleavage of a short hairpin RNA (shRNA) into small double-stranded RNA (dsRNA) of 21–25 nucleotide lengths, or transfected into the cell. The transfected siRNA is phosphorylated at 5′-ends by an endogenous kinase. The 5′-phosphorylated siRNA is incorporated into RNA-induced silencing complex (RISC) and unfolded. The antisense strand targets the RISC to homologous mRNA (sequence complementary to the siRNA guide) which is then cleaved by an endonuclease in the RISC complex (termed Slicer). The mRNA initially cleaved by Slicer is degraded by exonucleases and thus silenced

specific posttranscriptional regulators of gene expression. They are incorporated into related RNA-induced silencing complexes (RISCs) termed siRISC and miRISC respectively. RNAi is a naturally occurring post-transcriptional gene silencing mechanism in which 21–25 nucleotide (nt) double-stranded RNA (siRNA) with symmetric 2–3 nucleotide overhangs and 5′-phosphate groups prevent translation of homologous mRNA sequences by targeting these transcripts for degradation. siRNA can be generated from short hairpin RNA (shRNA) by the action of Dicer RNase. The transfected siRNA is, however, converted into active siRNA (5′-phosphorylation) by cellular kinase. The phosphorylated siRNA is incorporated into the RISC in an ATP-dependent process and unfolded. The antisense strand of siRNA targets the RISC to homologous mRNA, which is cleaved/degraded in gene silencing (Meister and Tuschl, 2004).

MicroRNA (miRNA) are single-stranded RNA, approximately 21–23 nt long. These small RNAs are derived from long primary transcripts termed pri-miRNAs, which are cleaved by the RNase III enzyme into precursor miRNA (pre-miRNA) hairpins of about 70 nt with a 3′ overhang of 2 nt. miRNA is excised from pre-miRNA by Dicer. The single-stranded miRNA guided with imperfect sequence complementarity to a target mRNA usually in the 3′ untranslated region (3′ UTR), induces translational repression whilst miRNA with full sequence complementarity can induce mRNA degradation (He and Hannon, 2004). Small RNAs, siRNA and miRNA may regulate expression of mRNA by the overlap pathways. Although siRISC and miRISC are functionally interchangeable, siRNA and miRNA programmed RISCs have distinct targeting functions in cells. Many endogenous miRNAs are genetically programmed to regulate gene expression and therefore are important for the growth and development of an organism. By contrast, siRNA are produced from dsRNA that are often synthesized from viruses or repetitive sequences or endogenously activated transposons. Therefore siRNA has been proposed to function in:

1. antiviral defense;
2. silencing mRNA that is overproduced or translationally aborted; and
3. guarding the genome from disruption by transposons (Tang, 2005).

Databases for siRNA and miRNA are available respectively at siRNAdb (http://sirna.cgb.ki.se/) and microRNA Registry (http://www.sanger.ac.uk/Software/Rfam/mirna/).

13.4.4 Posttranscriptional processing/modification

The immediate products of eukaryotic transcription, primary transcripts, are often not functional entities. Many of them undergo specific alternations in order to acquire biological activity (Rio, 1992), such as:

a) nucleolysis of the terminal or internal nucleotide chains;
b) appending oligonucleotides to their 3′ or 5′ ends; and
c) modification of specific nucleosides.

The three major classes of RNAs are altered in different ways.

13.4.4.1 Messenger RNA processing/modification. Most primary mRNA transcripts in prokaryotes function in translation without further modification. Eukaryotic mRNA transcripts, however, undergo extensive posttranscriptional modifications.

1. *hnRNP*: Eukaryotic mRNA are synthesized as precursors known as pre-mRNA, which are closed associated with a great variety of proteins forming heterogeneous nuclear ribonucleoproteins (hnRNPs).

2. *Eukaryotic mRNA are capped*: Specific guanylyltransferase appends a 7-methyl-GTP cap to the initial nucleoside of the transcript via a 5′-5′ linkage. The cap defines the eukaryotic translational start site, may be at the transcript's leading nucleoside (cap-1), at its first two nucleosides (cap-2) or at neither of these positions (cap-0).

3. *Eukaryotic mRNAs have poly(A) tails*: Eukaryotic mRNAs are invariably monocitronic and have well-defined 3′-poly(A) tails of 20–50 nt. Poly(A) tail is specifically complexed by poly(A) binding protein (PABP), which organizes the mRNA into ribonucleoprotein particle and protect mRNA from degradation. Alternative poly(A) modifies the 3′UTR, influencing the tissue distribution of the transcripts (Beaudoing and Gautheret, 2001).

4. *Eukaryotic mRNA consists of alternating coding regions and noncoding regions*: The primary eukaryotic transcript is pre-mRNA, which contains noncoding intervening sequences (IVS's or introns) whose aggregate length is generally longer than the combined coding sequences (exons). Exons are spliced together to form mature mRNA. Constitutive splicing refers to the removal of every intron and splicing of every exon at specific splice sites to form the mature mRNA. The splicing of different splice sites or by omitting entire exons is called alternative splicing. The alternative splicing contributes to the formation of protein isoforms and increase protein diversity posttranscriptionally (Boue *et al.*, 2003). ASD (http://www.ebi.ac.uk/asd) is the resource site for the altenative splicing (Stamm *et al.*, 2006).

5. *Exons are spliced in a two-stage reaction mediated by snRNP*: The sequences which are necessary and sufficient to define a splice junction (exon-intron junction) are an invariant GU at the 5′ boundary and an invariant AG at the 3′ boundary of the intron. The consensus sequence (Padgett *et al.*, 1986) at the exon-intron junctions of eukaryotic pre-mRNA is shown (the subscripts indicate the percentage of pre-mRNA in which the specified base occurs):

$$5' \ldots \text{Exon} .. A_{77}\overset{\substack{5'\text{ splice site}\\ \downarrow}}{}G_{100}U_{100}A_{60}A_{74}G_{84} .. \text{Intron} .. (U/C)_{77\text{–}91}NC_{78}A_{100}G_{100}\overset{\substack{3'\text{ splice site}\\ \downarrow}}{}G_{55} .. \text{Exon} \ldots 3'$$

The intron excision occurs via two transesterification reactions (subsection 11.7.2):

(1) The formation of a 2′,5′-phosphodiester bond between an intron A residue and its 5′-terminal phosphate group forms a lariat structure (Figure 11.23) and release the 5′ exon.

(2) The new free 3′-OH of the 5′ exon forms a phosphodiester bond with the 5′-terminal phosphate of the 3′ exon to yield the spliced product.

The intron is eliminated in its lariat form and is subsequently degraded. The splicing process that generates functional transcripts is termed functional splicing.

Splicing takes place in a 50–60S small nuclear ribonucleoprotein (snRNP) particle called spliceosome consisting of a pre-mRNA, small nuclear RNAs (snRNAs) and a variety of pre-mRNA binding proteins, which ensures the proper excision of all introns and precise splicing of exons.

6. Alu-*containing exons are alternatively spliced*: *Alu* elements are short (about 300 nt), interspersed elements that amplify in primate genomes (composing about 10% of the human genome) through a process of retroposition. A typical *Alu* is a dimer built of two similar sequence elements (right and left arms) that are separated by a short A-rich linker. Parts of *Alu* elements can be inserted into mature mRNAs by way of splicing termed exonization. The process is facilitated by sequence motifs that resemble splice sites, which are found with the *Alu* sequence, i.e. proximal AG and distal AG (Lev-Maor *et al.*, 2003).

7. *Alternative splicing*: In addition to generating protein diversity leading to proteome expansion, the alternative splicing can also regulate gene expression by splicing transcripts into unproductive mRNAs targeted for degradation. Nonsense-mediated mRNA decay (NMD) in an RNA surveillance function that recognizes mRNAs containing premature termination codons (PTC^+ mRNAs) and targets the transcripts for destruction rather than translation into proteins (Maquat, 2004). Nonsense and frameshift mutations, errors in pre-mRNA processing and alternative splicing are among the sources of PTC^+ mRNAs. In mammals, a termination codon is recognized as premature if it is more than about 50 nucleotides upstream of the site of removal of an intron.

8. *mRNA is methylated at certain A residues*: In vertebrate pre-mRNA, ~0.1% of A residues are methylated at N6 with m^6A commonly occurring in the sequence RRm^6ACX (R is purines and X is any bases but rarely G).

9. *RNA editing* (*Grosjean and Benne, 1998*): Some mRNAs differ from their corresponding genes in regard to C → U and U → C alternations, insertion or deletion of U's and insertion of G's or C's. The process whereby the primary transcript is altered in such a manner is called RNA editing. For example, a site specific cytidine deaminase catalyzes a substitutional editing, whereby C → U. Trypanosomes deletes poly(U), catalyzed by a high molecular weight particle called editosome with an aid of guide RNA.

13.4.4.2 Ribosomal RNA processing/modification

1. *Prokaryotic pre-rRNA processing*: The *E. coli* primary rRNA transcript contains 16S rRNA, 1 or 2 tRNA, 23S rRNA and 5S rRNA and some with 1 or 2 more tRNA (in order from 5′ to 3′ ends). The trimming and processing of pre-rRNA to mature rRNA involves various RNases such as RNase III, RNase P, RNase E, RNase F and RNase M at different stages.

2. *Methylation of rRNA*: The 16S and 23S rRNA are methylated at a total of 24 specific nucleosides. The methylation uses SAM as a methyl donor to yield N^6, N^6-dimethyladenine and $O^{2\prime}$-methyribose residues, which may protect the adjacent phosphodiester bonds from degradation by RNases.

3. *Eukaryotic pre-rRNA processing*: The primary rRNA transcript is 45S RNA containing 18S, 5.8S and 28S rRNA separated by spacer sequences (in order from 5′ to 3′ ends). An initial methylation of ~110 sites yields $O^{2\prime}$-methyribose residues (~80% of methylated sites) and methylated bases (~20% of methylated sites) such as N^6, N^6-dimethyladenine and 2-methylguanine. Subsequent trimming and processing by RNases produces mature rRNA.

4. *Splicing of some eukaryotic pre-rRNA*: Only a few eukaryotic rRNA transcripts contain introns, which are excised and self-spliced (subsection 11.7.2).

13.4.4.3 Transfer RNA processing/modification

1. *Prokaryotic pre-tRNA processing*: The primary tRNA transcript from *E. coli* contains from one to as many as four or five tRNA species. Some are found in primary rRNA transcription. The excision and trimming of these tRNA sequences resemble those for rRNA, and are carried out by various RNases.

2. *Eukaryotic pre-tRNA processing*: Many eukaryotic tRNA transcripts contain small introns, which do not disrupt the cloverleaf structure and are excised. Eukaryotic tRNA transcripts lack the obligatory −CCA sequence at their 3′ end. This sequence is appended stepwise by the action of tRNA nucleotidyltransferase using CTP and ATP as substrates.

3. *Posttranscriptional modification*: All tRNA have a large fraction (up to 25%) of modified bases. Most common modification is methylation catalyzed by methyltransferases using SAM as a methyl donor to yield various methylated bases such as m^3C, m^1A, m^7G and dim^2G.

13.5 TRANSLATION AND PROTEIN BIOSYNTHESIS

13.5.1 Protein translation: Overview

Translation converts the genetic information embodied in the base sequence (codon) of mRNA into the amino acid sequence of a polypeptide chain on ribosomes. Protein biosynthesis (Arnstein and Cox, 1992; Lee and Lorsch, 2004; Moldave, 1981) begins with the activated amino acids (aminoacyl-tRNA) and is characterized by three distinct phases; initiation, elongation and termination.

13.5.1.1 Synthesis of aminoacyl-tRNA. Amino acids are activated to 3′-*O*-aminoacyl-tRNAs (aa-tRNAs), which are delivered to ribosomes to form peptide bonds. The activation of amino acids is catalyzed by 20 different aminoacyl-tRNA synthetases (aRS's), each catalyzes the ATP-dependent esterification of its specific amino acid to the cognate tRNA (Arnez and Moras, 1997). Thus aRS reaction serves to:

1. activate the amino acid so that it may react to form a peptide bond; and
2. bridge the information gap (or difference) between amino acids and nucleotides (codons).

The aRS reaction occurs in two steps, as shown in Figure 13.15 (subsection 11.4.6 for reaction mechanism):

1. activation of the amino acid via ATP-dependent formation of an aminoacyl adenylate; and
2. transfer of the aminoacyl group from the aminoacyl adenylate to a specific tRNA.

Each aminoacyl-tRNA synthetase is specific for one amino acid and one or more isoaccepting tRNA. All 20 aRS's can be categorized into one of the two classes and their characteristics are summarized in Table 13.7.

Transfer RNA consist of ~80 nucleotides in a single chain with a majority of bases hydrogen bonded to assume a secondary structure, which is represented as a cloverleaf with three loops (some have extra variable loops) and the stem where the 3′- and 5′-ends of the molecule meet. These structural segments are designated as:

- *TψC or T arm*: 5 bp stem ending in a loop that usually contains TψC where ψ is pseudouridine;
- *anticodon arm*: 5 bp stem ending in a loop that contains the triplet anticodon;
- *D arm*: 3 or 4 bp stem ending in a loop that frequently contains D where D is dihydrouridine;

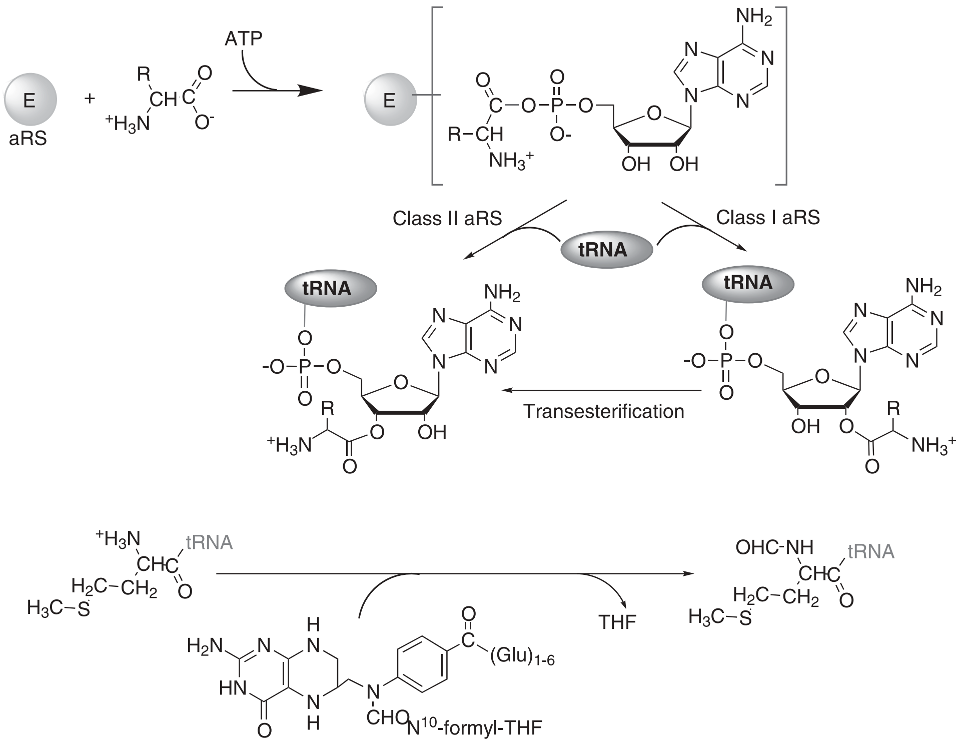


Figure 13.15 Reaction sequences for aminoacyl-tRNA synthesis. Synthesis of aminoacyl-tRNA catalyzed aminoacyl-tRNA syntheses (aRS) occurs in two steps, formation of enzyme-associated (activated) aminoacyl adenylate and transfer of the activated aminoacyl group to form esoteric linkage with either the 2′-OH (class I aRS) or 3′-OH (class II aRS) of the ribose on the terminal 3′–CCA of tRNA. The transesterification converts 2′-O-aminoacyl-tRNA to 3′-O-aminoacyl-tRNA which is the substrate for protein synthesis

- *extra variable loop* (for some tRNA);
- acceptor arm or stem: 7 bp stem including 5′-terminal nucleotide; and
- *3′-terminal sequence*: –CCA to which amino acid is appended in the aminoacyl-tRNA (aa-tRNA).

The anticodon can be found in the middle of the unpaired anticodon loop (7 nucleotides) of tRNA between positions 30 and 40 (at around position 35). The anitcodon is bordered by an unpaired pyrimidine, U on the 5′ side and often an unpaired alkylated purine on the 3′ side. Aminoacyl-tRNA synthetase must not only capable of discriminating between various tRNA, but also be able to recognize their cognate amino acids. The enzymes do not demonstrate a common set of rules that govern tRNA recognition nor limit their discriminatory features to the anticodon. Some recognition features include:

a) at lease one base in the anticodon (all amino acids except A,L,S);

b) one or more of the three bps in the acceptor stem (amino acids: A,D,G,H,M, Q,S,T,V,W); and

c) the base (discriminator base) at position 73 (the unpaired base preceding the 3′-CCA), which is invariant in the tRNA (all amino acids except T).

TABLE 13.7 Principal characteristics of the two classes of aminoacyl-tRNA synthetases

	Class I	Class II
Subunit structure	Mainly monomeric	Oligomeric usually homodimer
Conserved motifs:		
Dimer interface		Motif 1: +G(F/Y)xx(V/L/I)xxPϕϕ
Substrate binding	HIGH	Motif 2: +ϕϕxϕxxxFRxE
	KMSKS	Motif 3: ϕGϕGϕGϕϕERϕϕϕ
Active site fold	α/β fold known as Rossmann dinucleotide-binding fold	Antiparallel β-sheet
ATP binding	Sandwiched between GH of HIGH: Two H's interact with triphosphate. K of MSK interacts with α- and γ-phosphates.	Cradled between motifs 2 and 3: Adenine base stacks over F. R's of motifs 2 and 3 interact with the α- and γ-phosphates respectively.
Bound ATP conformation	Extended	Bent
tRNA binding:		
Acceptor stem	Minor groove side	Major groove side
Variable loop	Facing solvent	Facing protein
Amino acid specificity	Predominantly specific for larger, hydrophobic amino acids	Predominantly specific for smaller amino acids
Amino acid binding	Bind to more open and relaxed binding pocket.	Rigid, E of motif 2 anchors α-amino group of amino acids
Aminoacylation site	2′-OH (ribose) of —CCA	3′-OH (ribose) of —CCA
Codon recognition	15 of 16 tRNAs whose codons have a central U (except UUU for Phe) are aminoacylated by class I enzymes.	TRNA's specific for all 16 codons with a central C are aminoacylated by class II enzymes.
Editing mechanism	Elimination of mistakenly activated amino acids	Unknown

Notes: 1. The symbols + and ϕ represent positively charged and hydrophobilic amino acid residues respectively.
2. PheRS is the only exception being the member of the class II aRS that attaches Phe to the 2′-OH of tRNA^{Phe}.

13.5.1.2 Mechanics of protein translation. Protein translation begins with the activated amino acids and proceeds in three stages; initiation, elongation and termination. At each stage, the driving energy is provided by GTP hydrolysis and specific soluble protein factors participate in the processes which take place on ribosomes (Ramakrishnan, 2002):

1. Initiation involves binding of mRNA to the small ribosomal subunit, followed by association of initiator aminoacyl-tRNA that recognizes the first codon.
2. Elongation (polymerization) involves the synthesis of peptide bonds from N-terminus to C-terminus in which successive aa-tRNA adds one amino acid at a time to a growing peptide chain. In ribosome-mRNA complex, the ribosome reads mRNA in the 5′ → 3′ direction, while providing A-site or acceptor site (occupied by the incoming aa-tRNA), P-site or peptidyl site (occupied by the peptidyl-tRNA) and E-site or exit site (occupied transiently by the deacylated or outgoing tRNA) for the assembly of peptide chain. The peptide chain grows by transferring the peptidyl group to the incoming aminoacyl-tRNA to form a peptidyl-tRNA with one more residue.
3. Termination is triggered when the ribosome reaches a stop codon on the mRNA at which the polypeptide chain is released and ribosomal subunits dissociate from mRNA.

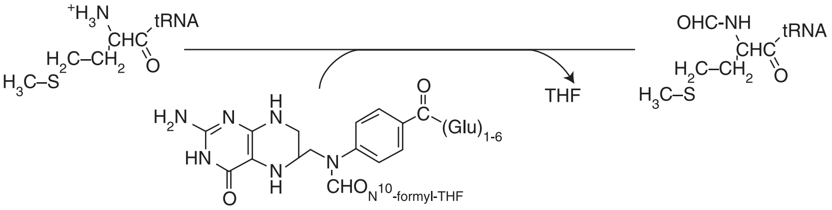
13.5.2 Protein translation: Processes

13.5.2.1 Chain initiation. Initiation of protein synthesis in prokaryotes involves binding of mRNA by small ribosomal subunit (30S), followed by association of the fMet-$tRNA_f^{Met}$ (initiator, *N*-formylmethionyl-$tRNA_f^{Met}$) that recognizes the initiation codon. A large ribosomal subunit (50S) then joins to form the 70S initiation complex. The tRNA ($tRNA_f^{met}$) that recognizes the initiation codon differs from the tRNA ($tRNA_m^{Met}$) that carries internal Met, although they both are aminoacylated by the same aRS (methionyl-tRNA synthetase) and recognize the same codon (AUG). Several structural features distinguish *E. coli* $tRNA_f^{Met}$:

a) the 5′-terminal base in the acceptor stem is not matched by a complementary base and therefore not hydrogen bonded;

b) a unique CCU sequence in its D arm; and

c) an exclusive set of three G:C bps in its anticodon arm. $tRNA_f^{Met}$ is first methionylated by methionyl-tRNA synthetase to Met-$tRNA_f^{Met}$, which is then converted to fMet-$tRNA_f^{Met}$ formyl transferase:



The N-terminal residue, fMet is either removed during the synthesis or deformylated post-translationally.

The selection of the proper initiation codon is aided by the base pairing between a pyrimidine-rich sequence at the 3′-end of 16S rRNA and complementary purine-rich tracts of 3–10 nt (known as Shine-Dalgarno sequence) at the 5′-end (centering ~10 nt upstream from the initiation codon) of prokaryotic mRNA. The assembly of the initiation complex and initiation of synthesis also require the participation of various soluble initiation factors (IFs, Table 13.8) in the initiation events:

TABLE 13.8 Protein factors of *E. coli* translation

Factor	Mol mass (kDa)	Function
Initiation factors:		
IF-1	9	Assisting IF-3 binding
IF-2	97	Binding of initiator tRNA and GTP
IF-3	22	Binding/release of 30S subunit and directing mRNA binding
Elongation factors:		
EF-Tu	43	Binding of aminoacyl-tRNA and GTP
EF-Ts	74	Displacing GDP from EF-Tu
EF-G	77	Binding of GTP and promoting translocation of ribosome
Release factors:		
RF-1	36	Recognition of UAA and UAG stop codons
RF-2	38	Recognition of UAA and UGA stop codons
RF-3	46	Binding of GTP and stimulation of RF-1/RF-2 binding

1. *Dissociation of 30S ribosome subunit*: IF-3 binds to the 30S subunit to promote the dissociation of the inactive 70S ribosomes, while IF-1 assists this binding and therefore facilitates the rate of ribosome dissociation.
2. *30S initiation complex*: The 30S subunit (assisted by IF-3) is complexed with the mRNA such that the initiation codon is aligned within the P-site. IF-2 (GTP) delivers the initiator fMet-tRNA$_f^{Met}$ to the P-site to form the 30S initiation complex.
3. *70S initiation complex*: IF-3 is released and the joining of 50S to the 30S initiation complex is accompanied by the IF-2 mediated GTP hydrolysis, which results in the release of IF-1 and IF-2 and the formation of the 70S initiation complex. Initiation results in the formation of the 70S initiation complex, fMet-tRNA$_f^{Met}$·mRNA·ribosome complex in which the fMet-tRNA$_f^{Met}$ occupies the P-site while the A-site is poised to accept an incoming aa-tRNA.

In eukaryotes, the initiator aminoacyl-tRNA is not formylated but instead is a unique tRNA functioning only in initiation as Met-tRNA$_i^{Met}$. Eukaryotic initiation requires a set of eukaryotic initiation factors (eIFs), as listed in Table 13.9, and proceeds in three steps:

1. *43S pre-initiation complex*: The initiator Met-tRNA$_i^{Met}$ is delivered as eIF2·GTP·Met-tRNA$_I^{Met}$ to the eIF1A and eIF3 bound 40S ribosomal subunit to form 43S pre-initiation complex consisting of Met-tRNA$_I^{Met}$, initiation factor and 40S ribosomal subunit. Binding of initiator Met-tRNA$_i^{Met}$ by the eukaryotic ribosomes is catalyzed by eIF4A in the absence of mRNA.
2. *48S pre-initiation complex*: The 43S pre-initiation complex binds mRNA at its 5′-terminal 7-methyl-GTP cap via eIF4F (mRNA cap-binding protein, eIF4E in association with eIF4G). The 5′-terminal 7-methyl-GTP cap and the 3′-polyA tail act synergistically to increase translational efficiency. eIF4G serves as a bridge between the cap-binding eIF4E, the polyA tail and the 40S subunit via eIF3. These interactions initiate the search of the initiation codon (AUG) by the 40S ribosomal subunit.
3. *80S initiation complex*: As 48S pre-initiation complex stops at the initiation codon, and hydrolysis of GTP in the eIF2·GTP·Met-tRNA$_I^{Met}$ ejects the initiation factors bound to the 40S ribosomal subunit. Release of these initiation factors allows the

TABLE 13.9 Eukaryotic initiation factors for protein translation

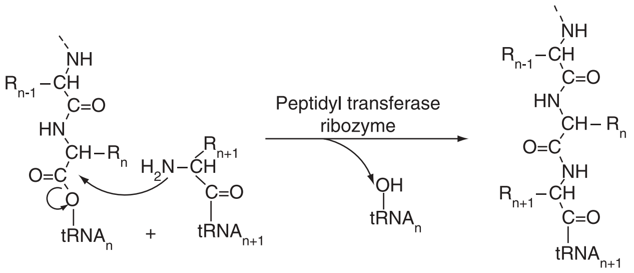
Factor	Mol mass (kDa)	Function
eIF1	15	Enhancement of initiation complex formation
eIF1A	17	Stabilization of Met-tRNA$_i^{Met}$ binding to 40S subunit
eIF2	125	GTP-dependent binding of Met-tRNA$_i^{Met}$ to 40S subunit
eIF2B	270	Promotion of G-nt exchange on eIF2
eIF2C	94	Stabilization of ternary complex in the presence of RNA
eIF3	550	Facilitation of Met-tRNA$_i^{Met}$ and mRNA binding
eIF4A	46	Binding of RNA and RNA helicase, promotion of mRNA binding to 40S subunit
eIF4B	80	Binding of mRNA, enhancing RNA helicase activity, assisting mRNA binding to 40S subunit
eIF4E	25	Binding of mRNA caps
eIF4G	153.4	Binding of eIF4A, eIF4E and eIF3
eIF5	48.9	Promotion of GTPase activity of eIF2, ejection of eIFs
eIF6		Dissociation of 80S by binding to 60S subunit

association of 60S ribosomal subunit with the 48S pre-initiation complex to form the 80S initiation complex and commences the translation.

13.5.2.2 Chain elongation. Elongation includes the synthesis of all peptide bonds of a polypeptide chain. This is accomplished, with the assistance of a set of protein elongation factors, by a repetitive cycle of events in which successive aa-tRNA adds to the A site and the growing peptidyl-tRNA occupying the P site of the mRNA·ribosome complex.

1. *Aminoacyl-tRNA binding*: The formation of aminoacyl-tRNA·EF-Tu·GTP, which binds to the A site as directed by codon. EF-Tu, is a member of the GTP-binding/hydrolysis proteins, which are often accompanied by a GTPase activating protein (GAP) and a guanine nucleotide releasing factor/protein (GRF/GNRP). In the case of EF-Tu, the ribosome is its GAP and EF-Ts is its GRF. Thus the GTP hydrolysis releases EF-Tu as EF-Tu·GDP and EF-Ts promotes the recycling of EF-Tu to EF-Tu·GTP by mediating the replacement of GDP with GTP. Only EF-Tu·GTP binds aminoacyl-tRNA but it does not bind fMet-$tRNA_f^{Met}$ (i.e. fMet-$tRNA_f^{Met}$ never reads the internal AUG codon).

2. *Transpeptidation*: In transpeptidation (peptidyl transfer), the peptide bond is formed via nucleophilic displacement of the P-site tRNA of the peptidyl-tRNA by the amino group of the aa-tRNA in the A-site. The nascent polypeptide chain is thereby lengthened at its C-terminus by one residue and transferred to the A-site tRNA (Figure 13.16). The peptidyl transfer reaction is catalyzed by the peptidyl transferase ribozyme of 23S rRNA of the 50S ribosome subunit.



3. *Translocation*: The elongation reaction transfers the peptide chain from the peptidyl-tRNA on the P-site to the aminoacyl-tRNA on the A-site by forming a new peptide bond. The E-site is transiently occupied by the deacylated tRNA as it exits the P-site. The new longer peptidyl-tRNA moves from the A-site into the P-site as the ribosome moves one codon further along the mRNA. This translocation vacates the A-site, which can accept the new incoming aa-tRNA. The translocation requires the binding of an elongation factor, EF-G·GTP to the ribosome.

The process of tRNA translocation occurs in two discrete steps (Figure 13.16):

1. Peptide bond formation shifts the acceptor end of the new peptidyl tRNA from the A-site to the P-site of the 50S subunit while its anticodon end remains associated with the A-site of the 30S subunit, a hybrid A/P binding state. The acceptor end of the deacylated tRNA moves concurrently from the P-site to the E-site of the 50S subunit whereas its anticodon end remains associated with the P-site of the 30S subunit in the hybrid P/E binding state (Green and Noller, 1997).
2. The binding of the EF-G·GTP to the ribosome causes the anitcodon ends of these tRNA together with their bound mRNA to move relative to the 30S subunit so that

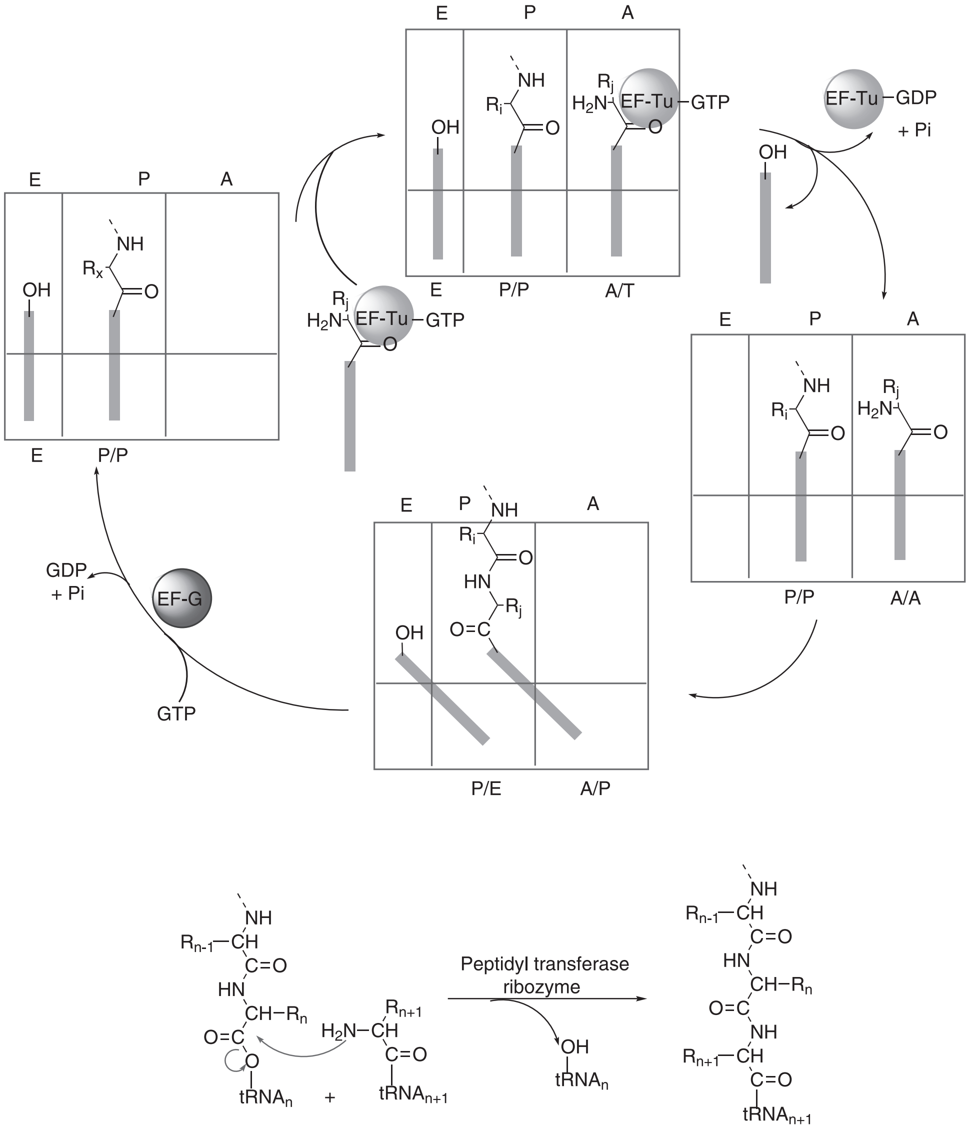

E
P
A
OH
NH
Rx
O
E
P/P
Ri
Rj
H2N
EF-Tu
GTP
A/T
EF-Tu
GDP
+ Pi
A/A
P/E
A/P
HN
O=C
EF-G
GDP
+ Pi
GTP
NH
Rn-1—CH
C=O
HN
CH—Rn
O=C
O
tRNAn
+
Rn+1
H2N—CH
C=O
tRNAn+1
Peptidyl transferase
ribozyme
OH
tRNAn
NH
Rn-1—CH
C=O
HN
CH—Rn
O=C
NH
Rn+1—CH
C=O
tRNAn+1

Figure 13.16 Diagram of elongation cycle and binding sites. The elongation cycle for protein biosynthesis is diagrammatically represented with tRNA binding sites according to hybrid state model which identifies six binding states. They are P/P (peptidyl-tRNA in the P site), A/T (aminoacyl-tRNA·EF-Tu·GTP binds in the A site), A/A (aminoacyl-tRNA in the A site), A/P (anticodon end of tRNA remains in the A site of the small subunit, but its peptidyl-ACC end occupies the P site of the large subunit), P/E (anticodon end of tRNA remains in the A site of the small subunit, but its deacylated —CCA end occupies the P site of the large subunit) and E (deacylated tRNA in the E site) states. Rectangular boxes represent ribosome subunits (upper large and lower small boxes for large and small subunits receptively) which are sub-divided into A, P and E sites. Thick elongated bars represent tRNA

the peptidyl-tRNA occupies the P-site of both ribosomal subunits in the P/P binding state. The deacylated tRNA occupies the E-site of the 50S subunit.

The elongation cycle is then completed by the release of the tRNA in the E-site and the binding of a new aminoacyl-tRNA in the A-site. The binding of tRNA to the A- and E-sites exhibits negative cooperativity. Upon binding an incoming aa-tRNA, the ribosome undergoes a conformational change that converts the A-site to a high affinity state and the E-site to a low affinity state, which consequently release the deacylated tRNA (pre-translocational state). Whereas at the end of an elongation cycle, the E-site binds the deacylated tRNA with high affinity, and the empty A-site has low affinity for aminoacyl-tRNA (posttranslocatonal state).

Eukaryotic peptide elongation follows similar processes as prokaryotes via aminoacyl-tRNA binding, transpeptidation and translocation, involving the A, P and E sites of the ribosome. Two elongation factors, EF1 and EF2 mediate the elongation steps. EF1 consists of two components, EF1A equivalent of EF-Tu and EF1B equivalent of EF-Ts. EF2 is the translocation factor that binds GTP and catalyzes hydrolysis of GTP that accompanies translocation.

13.5.2.3 Chain termination. Termination is triggered when the ribosome reaches a termination codon (UAA, UAG or UGA) on the mRNA. At this stage, the polypeptide chain is released and the ribosomal subunits dissociate from the mRNA. Three release factors, RF-1 and RF-2, recognize UAA/UAG and UAA/UGA respectively whilst RF-3, a GTP-binding protein, in complex with GTP stimulates the ribosomal binding of RF-1/-2. The binding of a release factor to the appropriate termination codon induces the peptidyl transferase ribozyme to water rather than to the aa-tRNA. The discharged tRNA dissociates form the ribosome and the release factors are set free with the concomitant hydrolysis of the RF-3 bound GTP. The resulting inactive ribosome releases its bound mRNA.

Termination in eukaryotes resembles that in prokaryotes but requires only a single release factor, eRF, which binds to the ribosome together with GTP. The hydrolysis of GTP triggers the dissociation of eRF form the ribosome.

13.5.3 Decoding mechanism

The transmission of genetic information requires accurate base pairing between nucleic acids. This includes translation of mRNA into protein by the ribosome on which accurate selection of aa-tRNA depends upon the correct pairing of three bases between the mRNA codon and the tRNA anticodon. For example, the codon-anticodon mismatches are efficiently rejected with error frequencies as low as 10^{-4}. Plausible mechanisms for the involvement of the ribosome in the decoding of mRNA codon:tRNA anticodon (Ogle *et al.*, 2003; Rodnina and Wintermeyer, 2001) include:

13.5.3.1 The tRNA selection pathway and kinetic proofreading. Aminoacyl-tRNA is delivered to the ribosomal A-site as a ternary complex with elongation factor EF-Tu and GTP. The events between the initial binding of the ternary complex and the incorporation of the amino acid into the peptide have been dissected into several kinetically distinguishable steps, as shown in Figure 13.17.

The tRNA anticodon associates reversibly (k_1 and k_{-1}) with the codon in the 30S A-site within an area referred to as the decoding center. In an initial selection (screening) step, noncognate aa-tRNA·EF-Tu·GTP complexes with no match between the anticodon

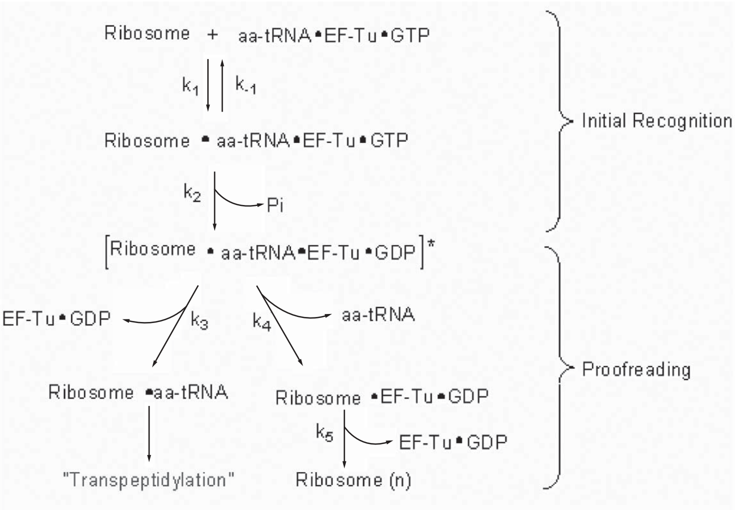


Figure 13.17 Kinetic events of aminoacyl peptidylation

and the codon, dissociate without stimulating GTP hydrolysis. Cognate (specific) codon–anticodon interactions stimulate a conformational change in EF-Tu leading to GTP hydrolysis (k_2). This intermediate complex can react in one of two ways:

1. EF-Tu·GDP dissociates from the ribosome with rate constant k_3 in the process termed accommodation. The accommodated aa-tRNA rapidly undergoes transpeptidation.
2. Aminoacyl-tRNA dissociates from the ribosome with rate constant k_4, thereby aborting the elongation step.

Subsequently, EF-Tu·GDP dissociates from the ribosome (k_5), allowing it to reinitiate elongation. In this second screening (k_4/k_3), known as the proofreading step, near-cognate aa-tRNA dissociates rather than being accommodated into the peptidyl transferase center on the 50S subunit. Rejection during proofreading is essentially irreversible because aa-tRNA can only efficiently enter the pathway as a ternary complex before GTP hydrolysis. The pathway is thus driven unidirectionally by GTP hydrolysis and ultimately by peptide-bond formation. It is recognized that in such an irreversible scheme, the overall accuracy could be as much as the product of the accuracy of the two rejection steps and thus enhance the specificity of codon–anticodon pairing. Furthermore, the energy derived from the binding of a cognate aa-tRNA anticodon induces conformational changes in the ribosome and is used to drive the irreversible chemical steps on the tRNA selection pathway.

13.5.3.2 Recognition of base-pair structure. The A-site of the 30S ribosomal subunit (the decoding center) is where the codon and anticodon pair is made up of four different domains: the head, shoulder, platform and helix 44. The 16S RNA bases of the decoding center specifically contact the cognate codon–anticodon pair by induced fit resulting in a closed conformation of the 30S subunit (rotating of the shoulder and the head domains toward the subunit center), compared to a more open structure of the unoccupied A-site. There are closed interactions between the three 16S RNA with the minor groove of the first two codon-anticodon bps that enhance the specificity of codon-anticodon recognition. The wobble base pairing at the third codon position allows some tRNA to recognize several different codons. This can occur because the contacts of the ribosome at the third codon position do not depend upon the precise shape of the minor groove, though the ribosome does impose certain geometric constraints (e.g. the overall width of the bp). Pairing at the wobble position is also heavily influenced by tRNA nucleoside modifications.

13.5.3.3 Linking base pair recognition, GTP hydrolysis and accommodation. The ribosome must use the energy of interaction with the cognate codon–anticodon duplex, not only to prevent aa-tRNA dissociation but also to switch on the GTPase activity of EF-Tu and to promote aa-tRNA accommodation into the peptidyl transferase site. The large distance between the sites of decoding and GTP hydrolysis suggests that the two must be coupled by conformational changes in either the ribosome or tRNA. Recent experiments suggest that cognate tRNA binding is found to induce a domain closure in the 30S subunit that is required for the subsequent steps in tRNA selection.

13.5.3.4 The role of 30S domain closure. An interaction between cognate aa-tRNA results in a transition from an open to a closed form of the 30S subunit. The influence of domain closure on tRNA selection begins prior to GTP hydrolysis and plays a role in the activation of EF-Tu. The dependence of GTP hydrolysis and aa-tRNA accommodation on domain closure would rationalize in structural terms for the discrimination of cognate tRNA versus near-cognate tRNA. The cognate tRNA bends into the decoding site, placing the anticodon in an orientation for contacts between its minor groove with the bases of 16S rRNA. It is conceivable that the cost of the closing movement could be increased by the presence of tRNA at the E-site (between the ‘head’ and the ‘platform’ domains), which could explain observations of reduced A-site affinity when the E-site is occupied by tRNA. Thus domain closure of the 30S subunit as a major rearrangement necessary for tRNA selection may rationalize the fidelity of tRNA selection.

13.5.3.5 Accommodation and rejection of aa-tRNA. The anticodon recognition by the decoding center favors an accommodated orientation of the anticodon stem–loop portion of tRNA. The bent tRNA conformation (energetical state conformation) observed in the aminoacyl-tRNA·EF-Tu·GTP complex is relaxed after dissociation of GDP from EF-Tu because the fully A-site-accommodated tRNA assumes the ‘classical’ L-shape (ground state conformation). It is conceivable that one of the forces driving accommodation is a thermodynamically downhill transition of the tRNA toward this ‘ground state’ conformation. The probability of the tRNA relaxing into the accommodated state without significant reorientation of the stem–loop portion of tRNA would thus be correlated with anticodon recognition by the decoding center. Thus the more restricted conformational space accessible to cognate aa-tRNA increases the probability (and hence the rate) of it reaching the accommodated state compared with near-cognate aa-tRNA. This influence of interactions at the decoding center on accommodation explains why cognate tRNA has a faster accommodation rate as well as a faster GTPase-activation rate.

13.5.4 Recoding, frameshifting and expanded genetic code

Messenger RNA is decoded in accordance with the assigned specificity and decoding mechanism involving tRNAs and rRNAs. In recoding, the rule changes are programmed in mRNA sequences via specific codons and other stimulator signals, which are manifested within mRNA folded structures (e.g. stem loops, pseudoknots) or through RNA-RNA interactions (Gesteland and Atkins, 1996). Four modes of the recoding are recognized:

1. *Codon redirection by frameshifting*: Frameshifting at a particular site allows expression of a protein from a mRNA with overlapping open reading frame. Both +1 and −1 programmed ribosomal frameshifting are possible. In the +1 frameshifting, the open reading frame switches to the +1 frame (Figure 13.18A) avoiding the stop codon and decoding the main portion of the message. In this case, a slippery codon, where the mRNA slides within the ribosome complex one nucleotide by breaking codon:anticodon pairing with peptidyl tRNA in the P-site and reestablishing pairing with an overlapping codon in the new frame.

Many retroviruses have overlapping *gag*, *pro* and *pol* genes such that one −1 frameshift event (Figure 13.18B) gives the gag-pro product and the gag-pro-pol product. The sites of frameshifting are heptanucleotide sequences of the form X XXY YYZ where mRNA slips one base with respect to the two tRNAs on the ribosome in A- and P-sites.

2. *Suppression and expanded genetic code*: Suppression (readthrough) is used to describe insertion of an amino acid for a stop codon. For example, the +1 frameshift at the slippery sequence acts as the four-base codon permitting an insertion of an amino acid and bypassing the stop codon (+1 frame suppressor). This tolerance of codons with more than three bases and enlarged anticodon loops of the translational machinery leads to genetic code expansion. Thus it is possible to design special tRNA as nonsense suppressors to UAG (amber stop codon) or +1 frame suppressors, which decode canonical N-base codon sequences with extended anticodon loops (N+4 nt). Combinatorial approach to the discovery of N-base suppressors indicates that the translational apparatus permits decoding of three-, four- or five-base codons and that each codon type favors tRNA of discrete sizes. Thus there are limits on both codon size (N = 3, 4, and 5) and tRNA anticodon loop size (N + 4) corresponding to codon length (Anderson *et al.*, 2002). The nonsense suppression on the amber stop codon using the expanded four/five-base codons can be employed to incorporate non-natural amino acids into proteins (Hohsaka and Sisido, 2002) via nonsense suppression mutagenesis (subsection 16.7.8).

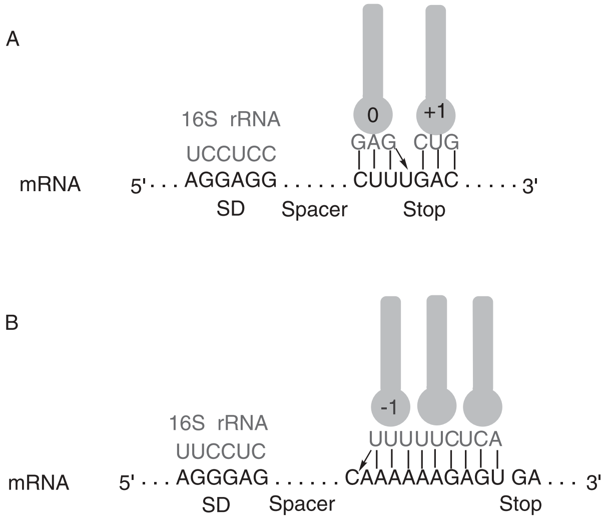

A
0
+1
16S rRNA
UCCUCC
GAG CUG
mRNA
5' . . . AGGAGG . . . . . . CUUUGAC . . . . . 3'
SD
Spacer
Stop
B
-1
16S rRNA
UUCCUC
UUUUUCUCA
mRNA
5' . . . AGGGAG . . . . . . CAAAAAAGAGU GA . . . 3'
SD
Spacer
Stop

Figure 13.18 Models of frameshifting. The models of +1 (A) and −1 (B) frameshifting with the Shine-Dalgarno sequence (SD paired with 16S rRNA), a spacer and the stop codon (UGA) which is bypassed by one nucleotide downstream or upstream respectively are shown

3. *Codon redefinition*: Recoding can reprogram the meaning of stop codons, thus redefinition competes with termination. The three stop codons, UAA, UAG and UGA have different efficiency. Furthermore, the first base 3′ to the stop codon has a large effect on the efficiency of termination with U>A>G>C in *E. coli* and A=G>>C=U in mammals. The meanings of code words are altered such that stop codons are redirected to encode selenocysteine (Sec) or Try at UGA and Gln at UAG. In prokaryotes, the mRNA enabling element that specifies a particular UGA as Sec codon is a stem loop structure immediately 3′ of the UAG that directs Sec-tRNA to the UGA. The stem loop sequence also redefines UAG and UAA. In mammals, the enabling element designed as SECIS is in the 3′ UTR of mRNA. The linear separation of UGA and the enabling element allows flexibility in UGA (it is read as a terminator if UGA is close enough to the SECIS).

4. *Translation over coding gap*: Disruptive sequences in mRNA are avoided via translational bypassing by ribosomes. The bypass requires matched codons where the peptidyl tRNA pairs interact first with one codon and then with the other by slipping of the mRNA. It appears to scan across the coding gap or fold into a structure that moves through the ribosome allowing an access to the matched codons (Weiss *et al.*, 1990).

13.5.5 Rescue system for stalled ribosomes

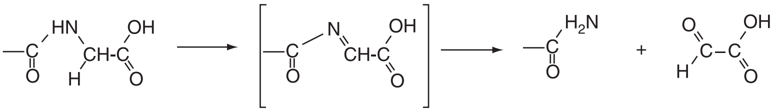
Truncated mRNAs lacking a stop codon cause the synthesis of incomplete peptide chains and stall translating ribosomes. In bacteria, a ribonucleoprotein complex composed of tmRNA (transfer-messenger RNA), a molecule that combines the functions of tRNA and mRNA, and small protein B (SmpB) rescues stalled ribosomes (Haebel *et al.*, 2004). The dual function of this molecule as both tRNA and mRNA facilitates a *trans*-translation reaction in which a ribosome can switch between translation of a truncated mRNA and the tag sequence of tmRNA.

Trans-translation also rescues ribosomes stalled at certain rare codons, possibly together with RNase to cleave mRNA in the ribosomal A-site. The *trans*-translation system is conserved among bacteria, some mitochondria and chloroplasts. The core component of the rescue system is a chimeric RNA molecule of 350–400 nucleotides termed tmRNA. The 5′- and 3′-ends of tmRNA for a tRNA-like domain with four extensions including the T(TψC)-arm, D-loop and the acceptor stem. The 5′- and 3′-end sequences of all tmRNAs can be folded into a tRNA-like structure of a 7-bp stem corresponding to an amino acid acceptor-arm and a 5-bp stem with a 7-base loop in the 3′-end regions corresponding to the TψC arm in tRNA. The consensus sequence of the TψC arm in tRNA, G TψCRA (R is A or G) is also present in all tmRNAs, which also the terminate with a CCA(3′) sequence as in tRNA. Psudoknot 1 (pk1) the mRNA-like domain, contains the internal open reading frame (ORF) encoding the approximately 10 amino acid degradation tag (C-terminal ANDENYALAA) and pk2–pk4 complete the secondary structure of tmRNA. Only pk1 is essential for *trans*-translation. A small approximately 160 amino acid protein component, SmpB, forms a tight complex with tmRNA and is required for its biological function by facilitating binding of tmRNA to the ribosome. *Trans*-translation ensues with the transfer of the incompletely synthesized polypeptide chain from P-site bound polypeptidyl-tRNA to A-site bound alanyl-tmRNA, during which the polypeptidyl-tmRNA is moved to the P-site and the tmRNA ORF replaces the defective mRNA. After several elongation cycles, the internal stop codon of tmRNA terminates translation, releasing the tagged protein, and the dissociation of ribosomal subunits. The degradation of defective mRNA might be facilitated by RNase R, a 3′,5′-exonuclease that associates with the SmpB-tmRNA complex (Withey and Friedman, 2003).

13.5.6 Posttranslational modifications of protein

A large number of proteins undergo post-translational modifications (Wold, 1981; Harding and Crabbe, 1992). Modifications of proteins post-translationally provide the means for targeting/translocation of many proteins and regulation of protein activities/functions. Table 13.10 summarizes common post-translational modifications that regulate enzyme activities.

Phosphorylation/dephosphorylation has been treated in the Chapters 11 and 12. Some other common post-translational modifications will be briefly described. In the amidation reaction, the terminal amine is derived from the C-terminal glycine via:



Glu (first 40 residues from the N-terminus) of certain proteins involved in blood clotting and bone structure, is carboxylated to γ-carboxyglutamate (Gla) by the vitamin K-dependent carboxylase. Gla is invariably involved in Ca^{2+}-binding. Hydroxylation of P and K occurs in procollagen in which P of —X-P-G— is γ-hydroxylated and K in —X-K-G— is δ-hydroxylated. These hydroxylations are essential to the maturation of collagen in the formation of triple helix (γ-OH-Pro) and cross-linkages (δ-OH-Lys).

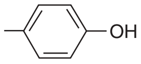
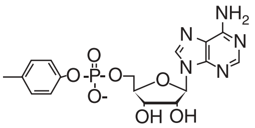
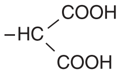
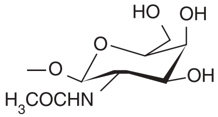
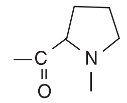
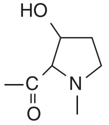
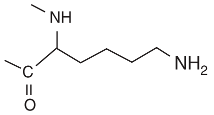
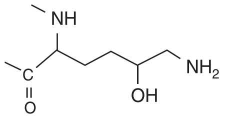
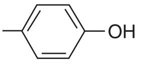
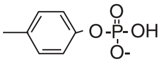
13.5.6.1 Acetylation and acylation. Acetylation of ε-amino groups on conserved Lys residues in the N-terminal domains of the core histones, H2A, H2B, H3 and H4, is the most common post-translational modification of chromatin. Acetylation substantially weakens the constraints on DNA imposed by the core histones and provides molecular mechanism by which DNA becomes accessible to transacting factors while maintaining a nucleosome architecture (Wade *et al.*, 1997). Histone acetylation is reversible and is controlled by a group of acetyltransferases and deacetylases. The balance between histone acetyltransferases and deacetylases determines the accessibility of the chromatin to the transcriptional machinery. Thus the acetylation status of histones is a key determinant of transcriptional activity. Transcription activators are often associated with histone acetyltransferases and repressors can interact with histone deacetylases (Ng and Bird, 2000).

Fatty acids (myristic and palmitic acids) are attached to a different sub-population for a limited number of eukaryotic cellular proteins in acylation reaction. Myristic acid is normally attached to proteins via an amide bond to an N-terminal Gly by myristoyl CoA:protein *N*-myristoyl transferase co-translationally whereas palmitic acid is attached to proteins via thioester bond to Cys catalyzed by S-acyl transferase. Myristate and palmitate act as membrane anchors and mediators of protein–protein interactions (Smotrys and Linder, 2004).

13.5.6.2 Methylation and prehenylation. In bacterial chemotaxis, the methyl group from *S*-adenosylmethione (SAM) is transferred to the γ-carboxyl of Glu residues. In eukaryotic cells, protein methylation takes place mainly at carboxyl groups, ε-amino of Lys (mono-, di- and tri-), ω-amino of Arg, or imidazole-*N* of His residues. Protein methylation catalyzed by methyl transferases, is involved in cellular stress responses, aging/repairing and dynamic regulatory events in ligand-stimulated signal transduction (Aletta *et al.*, 1998).

In relation to methylation, some proteins undergo prehenylation, which adds either farnesyl ($H[CH_2(CH_3)C{=}CHCH_2]_3$—) or geranylgeranyl ($H[CH_2(CH_3){=}CHCH_2]_4$—)

TABLE 13.10 Some common posttranslational modifications

Modification	Converter enzyme	Modifier	Residue modified	Product
Acylation	Acyl transferase	Acetyl CoA	K: ε—NH$_2$	—NH-COCH$_3$
		Myristoyl CoA	G: N-terminal	—NHCO(CH$_2$)$_{12}$CH$_3$
		Palmitoyl CoA	C: —SH	—SCO(CH$_2$)$_{14}$CH$_3$
Nucleotidylation (Adenylylation)	Adenylyl transferase ADP-ribosyl transferase	ATP NAD$^+$	Y: 	
Amidation	Glycine lyase	FAD	cG	—CONH$_2$ + OHCCOOH
Carboxylation	Carboxylase	CO$_2$, O$_2$, vit K	E: —CH$_2$COOH	 $-HC(COOH)_2$
Glycosylation	NDP:glycosyl transferase	NDP-glycoses	S/T: —OH C: —SH	
Hydroxylation (monooxygenation)	Prolylyl hydroxylase Lysyl hydroxylase	Fe^{2+}, O$_2$	P 	
			K 	
Isomerization (redox)	Disulfide isomerase	GSSG	C: —SH	—S-S—
Methylation	Methyl transferase	SAM	D/E: —COOH	—COOCH$_3$
			K/N: —NH$_2$	—NHCH$_3$/—N(CH$_3$)$_2$
Phosphorylation	Kinase	ATP/GTP	S: —OH	—O-PO$_3$H$^-$
			Y: 	
Proteolysis	Protease	H$_2$O	Varied	Peptides
Protein splicing	Extein/intein		Removal of inteins	Spliced exteins

Notes: 1. Abbreviations used are: one-letter codes for amino acid residues; c, C-terminal residue; n, N-terminal residue; GSSG, glutathione; NDP, guanosine/uridine diphosphonucleotides; SAM, S-adenosylmethionine; vit., vitamin.

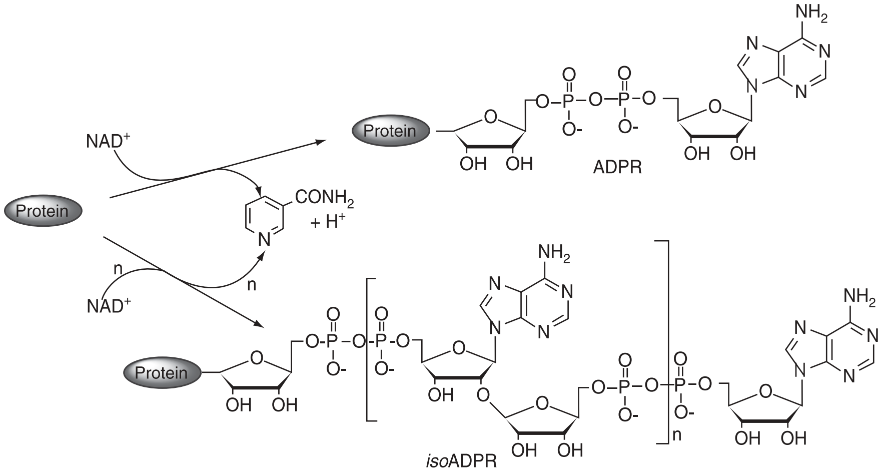
2. Only representative modified residues/products and modifiers are listed. For examples, methylation can take place, in addition to D, E and K, at imidazole of H, thioether of M, amide of N, and guanidinum of R. For glycosylation, only the products of the first monosaccharide units attached to Ser and Asn are shown. Nucleotidylation also uses UTP to give O-(5′-uridylyl)tyrosine. Hydroxylation of proline produces 3-hydroxyproline (as shown) and 4-hydroxyprolin (not shown).

3. The incorporation of covalently bound coenzymes, such as heme to C, FMN and FAD to thiol of C, and lipoate to ε-NH$_2$ of K may be considered as posttranslational modifications which are essential for catalytic activities of respective enzymes.

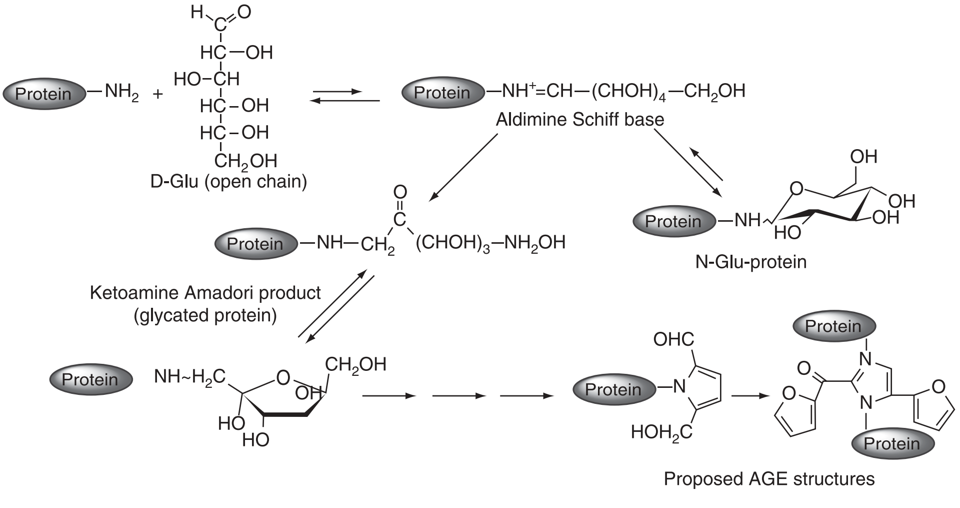
4. References cited are: McIlhinney, R.A.J. (1990) for acetylation; Kingdon *et al.* (1967); Bradbury and Smyth (1991) for amidation; Suttie (1980) for carboxylation; Gahmberg and Tolvanen (1996) for glycosylation; Gallop and Paz (1975) for hydroxylation; Freedman *et al.* (1994) for redox isomerization; Paik and Kim (1975) for methylation; Deutscher and Saier (1988) for phosphorylation; Steiner *et al.* (1992) for proteolysis; Cooper and Stevens (1995) for protein splicing.

isoprenoids covalently to the thiol of Cys residues at or near the C-terminal in the so-called CAAX motif (Zhang and Casey, 1996) of intracellular proteins. The C-terminal carboxyl group is then methylated. Prehenylation promotes membrane interactions of the modified proteins.

13.5.6.3 Mono- and poly-ADP-ribosylation. Mono-ADP-ribosylation and poly(ADP-ribose) modification catalyzed by ADP-ribosyl transferase and poly(ribose) synthetase respectively, append mono- and poly ADP-ribose (ADPR) moieties (Hayaishi and Ueda, 1977). In poly ADPR, the polymeric chain consists of *iso*ADP-ribose, 2′-(5″-phosphoribosyl)-5′-AMP.

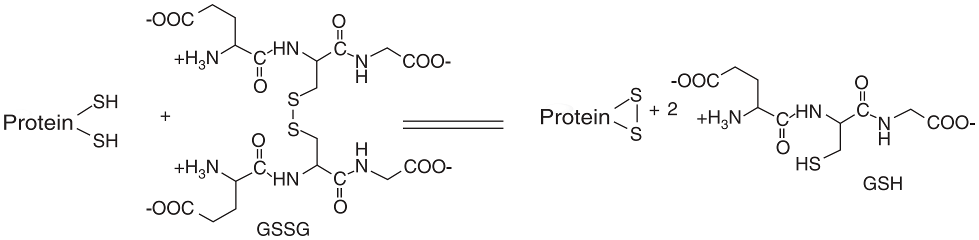

NAD+
Protein
CONH2
+ H+
n
NAD+
n
ADPR
NH2
isoADPR
n

13.5.6.4 Glycosylation versus glycation. Glycosylation of proteins is an important process for protein structures, targeting and translocation. The covalent attachment of saccharides to proteins can occur by enzymatic glycosylation, which incorporates monosaccharide, disaccharide and oligosaccharide chains (biosynthesis of glycoproteins has been described previously in section 13.1), and chemical glycation, which derivatizes α-amino of N-terminal residue or ε-amino of Lys with monosaccharides. The chemical glycation, which occurs spontaneously in proteins exposed to high concentration of reducing sugars, involves the formation of Schiff's base adduct, Amadori intermediate and cross-linked fluorescent product called advanced glycated end-product (AGE) (Ledl and Schleicher, 1990).


Protein
D-Glu (open chain)
Protein—NH+=CH—(CHOH)4—CH2OH
Aldimine Schiff base
Protein—NH—CH2—C(=O)—(CHOH)3—NH2OH
Ketoamine Amadori product
(glycated protein)
N-Glu-protein
Proposed AGE structures

The enzymatic glycosylations are catalyzed by NDP-glycosyltranferases using nucleoside diphosphates (NDP) such as GDT and UDP as glycosyl carriers. Monosaccharide and disaccharide units form *O*-glycosidic linkages with Ser, Thr, hydroxyPro, hydroxyLys, such as *O*-[4-*O*-(β-galctosyl)-β-xylosyl]Ser, *O*-[3-*O*-(β-glucosyl)-α-fucosyl]Thr, *O*-(glucosylarabinosyl]hydroxyPro, *O*-[2-(α-glucosyl)β-galactosyl]hydroxyLys, and *S*-glycosidic linkages with Cys (e.g. *S*-digalactosylCys and *S*-triglucosylCys).

13.5.6.5 Thiol-disulfide isomerization. Disulfide bond formation between Cys residues takes place commonly in the ER during protein biosynthesis. The thiol-disulfide isomerization can occur by either chemical exchange reaction or enzymatic reaction catalyzed by protein disulfide isomerase via mixed disulfide intermediate (protein–S-SG).



The disulfide formation depends on the protein conformation that places Cys residues into appropriate proximity and the disulfide redox potential that determines the intrinsic stability of protein disulfide bonds. For catalytic activity, the reduced dithiol form of protein disulfide isomerase is required. Protein disulfide isomerase is a folding catalyst that assists protein folding (Gilbert, 1997). The enzyme increases the rate of the overall folding process of the substrate protein without altering its pathway.

13.5.6.6 Proteolytic cleavage. Proteolytic cleavage is one of the most prevalent types of protein posttranslational modifications. Proteolytic modification of proteins is catalyzed by various endo- and exo-peptidases (subsection 12.8.1 for catalytic mechanism) and play essential functions in:

1. introduction of structural/functional diversity to proteins;
2. activation of pro-proteins (subsection 11.5.2 for zymogen activation); and
3. sorting and targeting of proteins (subsection 13.5.5).

Most prohormones and neuroendocrine precursors are processed at K-R or R-R sequences, while the precursors of many growth factors and various plasma proteins have complex multibasic cleavage sites of the general sequence, R-X-K/R-R (Steiner *et al.*, 1992).

13.5.6.7 Protein splicing. Protein splicing involves the precise excision of an intervening protein sequence (intein) from a precursor protein, coupled to peptide bond formation (ligation) between the flanking amino-terminal (N-extein) and carboxyl-terminal (C-extein) segments to give a spliced protein product (Cooper and Stevens, 1995; Paulus, 2000). The splice junctions of all inteins are closely related (Perler *et al.*, 1997):

↓ ↓
n————————Cys/Ser———— HisAsnCys/Ser/Thr————c
N-Extein Intein C-Extein

Asn and His are the respective C-terminal and penultimate residues of the intein. Cys, Ser and Thr are the residues immediately C-terminal to each of the two splice junctions (i.e. N-terminal residues of intein and C-extein). The information concerning intein sequences, properties and references are available online at InBase (http://www.neb.com/neb/inteins.html. The proposed splicing mechanisms are shown in Figure 13.19.

13.5.7 Protein translocation

In eukaryotes, most proteins destined for secretion, deposition in certain organelles or integration into the membrane are transported across the endoplasmic reticulum (ER) mem-

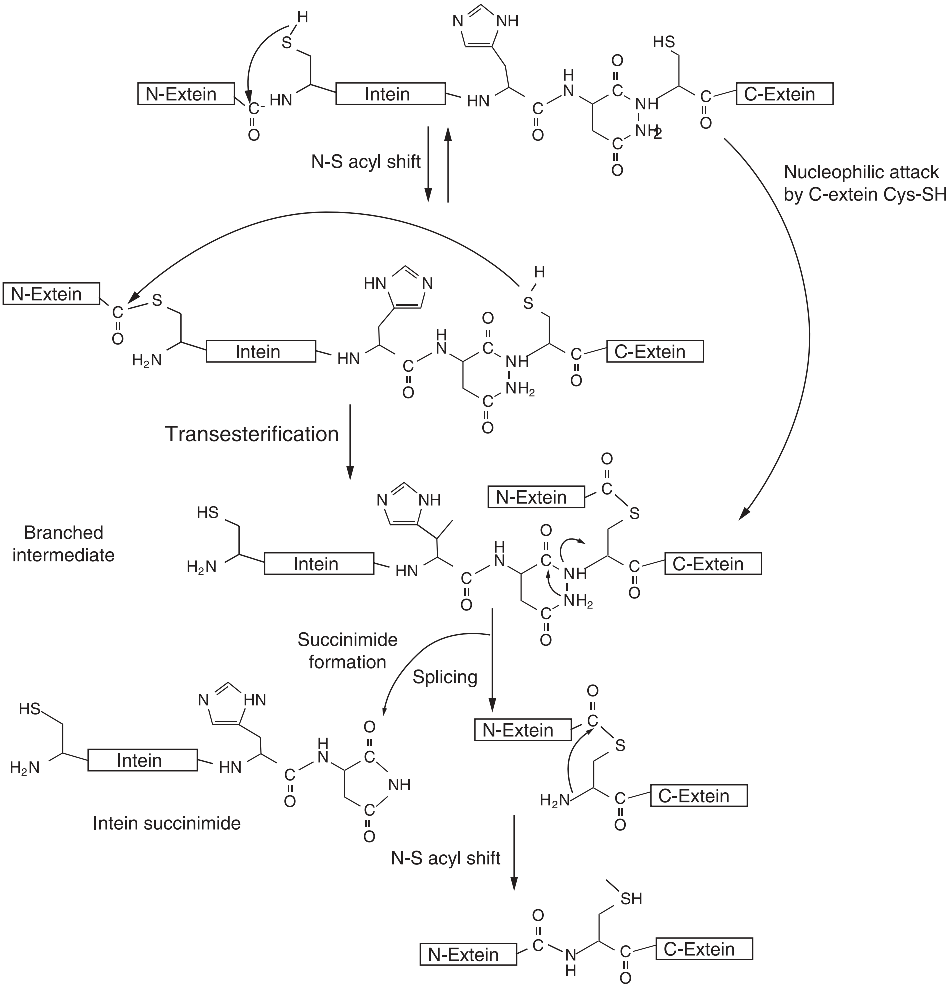

N-Extein
Intein
C-Extein
N-S acyl shift
Nucleophilic attack by C-extein Cys-SH
Transesterification
Branched intermediate
Succinimide formation
Splicing
Intein succinimide
N-S acyl shift

Figure 13.19 Proposed protein splicing mechanisms. Two possible mechanisms for protein splicing of splice junctions with Cys as the N-terminal residues of intein and C-extein. The initial step involves either N-S shift (or N-O shift for Ser as the N-terminal residue of intein) at the N-intein junction or nucleophilic attack by the N-terminal residue of C-extein. The function of the penultimate His is not explicitly indicated, however it may acts as general base-acid to assist this and subsequent step. Both mechanisms involve the formation of the branched intermediate which undergoes splicing to remove the intein as the succinimide derivative and the extein product after the N-S (or N-O) shift

brane. The ER is the site at which secretory proteins and membrane proteins enter and are distributed. The processes of protein translocation into the ER (across or integrate) are characterized by some common features (Corsi and Schekman, 1996; Rapoport *et al.*, 1996; Dalbey and von Heijine, 2002):

1. Proteins targeted for translocation are synthesized in a precursor form carrying an N-terminal stretch of amino acid residues (leader peptide) that serves as a signal sequence.
2. Two translocation pathways are operative in eukaryotes, i.e. co-translational and post-translational. In the co-translational pathway, transport occurs while the polypeptide chain is being synthesized on a membrane-bound ribosome. In the post-translational pathway, the polypeptide chain is completed in the cytoplasm before being transported. In prokaryotic, ribosomes do not seem to be tightly bound to the membrane and most proteins may be transported post-translationally or after much of the chain has been synthesized.
3. a) The co-translational process is initiated by binding of the signal recognition particle (SRP) to the signal sequence after it has emerged from the ribosome. SRP binds to the Met-rich signal sequence through its M-domain while its G-domain binds GTP.
   b) The fidelity of co-translational translocation is enhanced by the two consecutive signal-sequence recognition events, i.e. one in the cytosol by the SRP and one in the membrane.
   c) The selectivity of SRP-signal sequence interaction probably involves nascent polypeptide-associated complex (NAC), which is involved in maintaining the fidelity of co-translational precursor targeting to the translocon. NAC serves as an inhibitor for the SRP-independent interaction of the ribosome with the ER membrane and interaction of SRP with nascent chains lacking signal sequences.
   d) The target complex (consisting of ribosome, the nascent polypeptide chain and SRP) formation involves the binding of SRP-GTP to SRP receptors (SR, docking proteins) and the ribosome to ER membrane proteins.
4. In the post-translational translocation, the completed polypeptide chain is presented to the ER membrane in a complex (e.g. Sec complex) with cytosolic proteins, particularly with chaperones, and these must be released before translocation can occur. SRP is not involved in the post-translational translocation.
5. Polypeptide chains are transported through the membrane at specific translocation sites called translocons. Translocon consists of a protein complex catalyzing energy transduction, post-translational modifications and movement of the translocating protein across the membrane.
6. Polypeptide chains are transported across the membrane through an aqueous channel formed from membrane proteins. This protein-conducting channel is transiently formed so that a polypeptide chain can move across the pho-spholipid bilayer. The protein-conducting channel exists for both co- and post-translational processes and opens in two dimensions, i.e. perpendicular to the membrane to let hydrophilic polypeptide chains cross, and within the plane of the membrane to let hydrophobic anchors of membrane proteins into the phospholipid bilayer.

7. Proproteins interact with molecular chaperones in loosely-folded, translocation-competent conformations, particularly for a polypeptide chain targeted for the post-translational process.
8. The signal sequence is proteolytically cleaved by signal pepetidase (member enzyme of translocon) once protein is translocated. The eukaryotic signal peptidases have multiple catalytic subunits with broad substrate specificity. The enzyme generally cleaves at a site that has small aliphatic residues at position −1 and −3. On-line prediction of signal peptide and cleavage site is available at PrediSi (http://www.predisi.de).

The eukaryotic co-translational and post-translational translocations are diagrammatically compared in Figure 13.20.

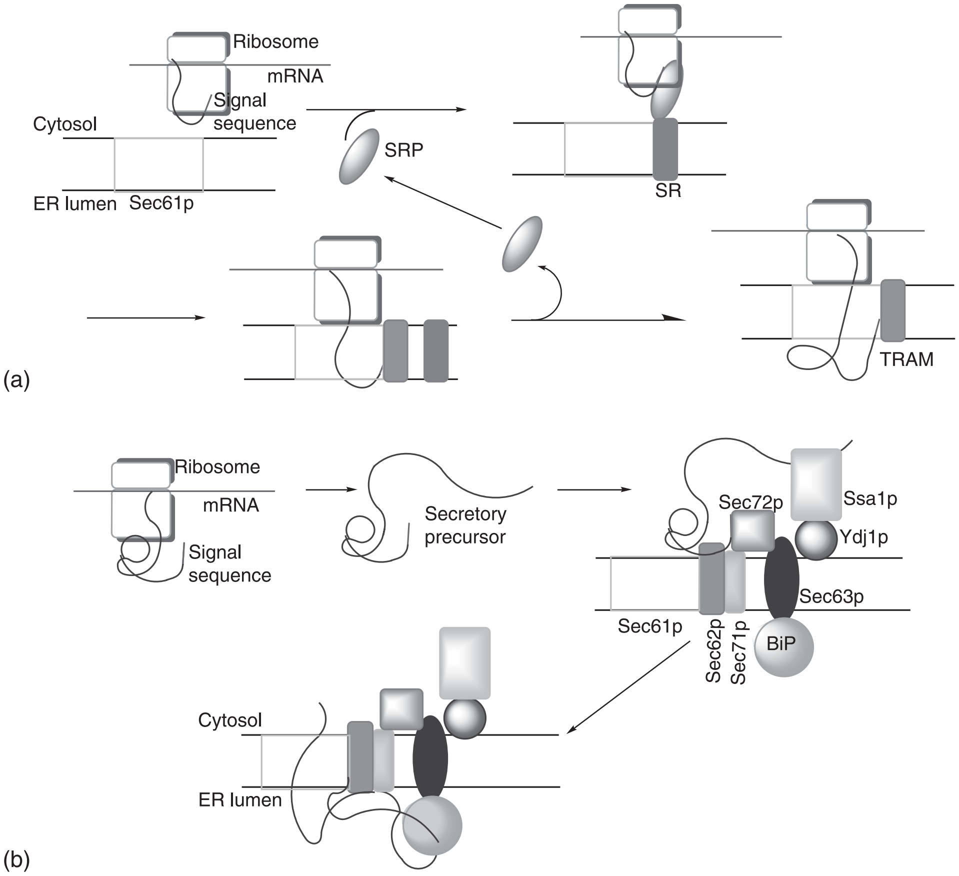

Ribosome
mRNA
Signal sequence
Cytosol
ER lumen
Sec61p
SRP
SR
TRAM
(a)
Ribosome
mRNA
Signal sequence
Secretory precursor
Sec72p
Ssa1p
Ydj1p
Sec63p
Sec61p
Sec62p
Sec71p
BiP
Cytosol
ER lumen
(b)

Figure 13.20 Diagrams for eukaryotic translocations across the ER membrane. The mammalian co-tanslational translocation (a) and yeast post-translational translocation (b) of polypeptide chain are diagrammatically represented. Abbreviations used are: SRP, signal recognition particle; SR, SRP receptor and TRAM, translocating chain-associated membrane protein. Sec61p spans the ER membrane multiple times and likely forms the translocation channel. The cytosolic components, Ssa1P and Ydj1p which maintain the nascent polypeptide chain in the unfolded state in the post-translational translocation. The nascent polypeptide-associated complex (NAC) which maintains the fidelity of co-translational precursor and the role of GTP are not shown

13.6 FOLDING OF BIOMACROMOLECULES

13.6.1 Overview

Both RNA and protein can adopt complex globular structures, i.e. both fold into characteristic 3D structures. The folding of RNA and protein represents information transfer from one dimension to three dimensions. Thus the fundamental issue in both protein and RNA folding is how a linear sequence of residues dictates three-dimensional structure capable of biological functions. Several characteristic differences seem to dictate different folding behaviors for the two types of biomacromolecules.

1. *Strength of interactions*: Nucleotides have the potential for far stronger interactions than amino acids. An additional bp to a secondary structural helix contributes –4.18 to –2.55 kJ/mol. Isolated protein secondary structures are usually unstable in the absence of the rest of the protein. A folded protein has only a modest stability in the order of –41.84 kJ/mol.

2. *Complementarity of interactions*: RNA structures depend heavily on complementary pairs of hydrogen-bond acceptors and donors for their secondary and tertiary structures. Protein secondary structures depend on complementary alignments of the backbone amide and carbonyl groups for hydrogen bonding. Whereas the hydrophobic bonding and electrostatic interactions, which contributes to protein tertiary structures, do not rely on complementarity of the side chains.

3. *Side chain placement*: In proteins, the α-helix and β-sheet places the chemical groups of amino acid side chains on the surface of the secondary structures, optimally positioned for tertiary structural interactions, whereas the secondary structure of RNA places the chemical groups of nucleotides in the interior of the A-form duplex largely inaccessible for tertiary structure formation.

4. *Specific ion association*: The stability of RNA depends significantly on the status of specific ion association, while that of protein is less important.

13.6.2 RNA folding

Ribonucleic acids often fold into structures with unique and irregular shapes in physiological environments (37°C, 150 mM Na^+ and 10 mM Mg^{2+}). Intrinsic events during RNA folding include conformational search and metal ion binding. RNA forms regular secondary (helical) structures from canonical base pairing (Watson–Crick pairing) and additional noncanonical and tertiary interactions presumably account for the diversity of RNA folds. Three G·C bps suffice to close a loop of 4, 5 or 6 nucleotides. More A·U pairs are needed for loop closure, but a random-sequence RNA will have approximately half of its bases paired (Onoa and Tinoco, 2004). Since biomacromolecules undergo unfolding and refolding at equilibrium, an understanding of RNA folding may be gained from RNA unfolding studies (Draper, 1996). In general, RNAs unfold their tertiary structure first before the secondary structure.

13.6.2.1 tRNA. Transfer RNA such as $tRNA_f^{met}$ unfolds in a few discrete steps with an elevated temperature. A set of tertiary interactions unfolds first to be followed by the unfolding of the cloverleaf secondary structure in a series of additional steps. The tertiary structure refers to cooperatively folding domains of RNA and therefore those interactions that cross-link individual segments of RNA structure are considered as tertiary interactions. The D arm unfolds at the same time as the tertiary structure and unfolding of the TψC arm, anticodon arm and acceptor arm follow in that order.

13.6.2.2 rRNA. A 58-nucleotide fragment derived from the large subunit RNA (1051–1108 rRNA) consists of A1 hairpin connected via A2 helix to B and C stems. An invariant base, A1078 of the A2 helix appears to play a dual role in pairing with U1061 and participating in the tertiary interaction. Specific ion coordination is important to the stability of rRNA structure. As the temperature is raised, the rRNA fragment unfolds its tertiary interactions prior to any of the secondary structures, which unfold in the order, A2 helix, B/C stems and A1 hairpin.

13.6.2.3 mRNA. The α mRNA fragment, which serves as the target site for a translational repressor (ribosomal protein S4), consists of four helical segments, I, II, III and IV from 5′-end to 3′-end to form a nested pseudoknot. In the temperature-dependent melting, the unfolding takes place in the order of tertiary interaction, helix IV, helix I and helices II + III.

13.6.3 *In vitro* protein folding pathway

The information for folding each protein into its unique 3D structure resides within its amino acid sequence. For folding to proceed, the sum of favorable interactions (e.g. hydrophobic interactions, hydrogen bonds, electrostatic interactions, van der Waal's interactions) must exceed the unfavorable loss in configurational entropy (~4.2–8.4 kJ/mol destabilizing at room temperature) (D'Aquino *et al.*, 1996). It appears that proteins are constructed from a small catalog of recognizable parts that fit together in a limited number of ways (Fitzkee *et al.*, 2005; Shakhnovich *et al.*, 2001).

In vitro studies of protein folding reactions begin with the protein in solution conditions designed to disrupt the noncovalent interactions that stabilize the native conformation. The unfolding of proteins is normally accomplished by thermal, pH and/or chaotropic agents (e.g. guanidine hydrochloride, urea). Refolding of most proteins is a self-assembly process that occurs spontaneously under the appropriate conditions in 10^{-1}–10^{3} s. This suggests that proteins do not fold by sampling all possible conformations randomly until the one with the lowest free energy is encountered. Instead the folding process must be directed in some way. The folding rate of a protein, to a first approximation, can be related to the entropic cost of forming the native-like topology. The information needed for folding is distributed throughout the polypeptide chain in a network of diverse interactions, each contributing to the outcome. Protein folding via random and unbiased searching of all possible conformations is unlikely. There appear to be pathways that simplify the mechanism of folding by breaking it down into sequential steps.

13.6.3.1 Folding pathway exemplified. Hen's egg white lysozyme is a monomeric protein of 129 amino acid residues. It has four α-helices, which make up one domain (α-domain) encompassing the N- and C-terminal segments of the protein, and triple-stranded antiparallel β-sheets with a long loop make up a second domain (β-domain). In addition, there are two 3_{10} helices, one in each domain, and a short region of double stranded antiparallel β-sheets that link the two domains. Four disulfide bridges, which are maintained during the folding process, stabilize the structure. Two distinct types of folding intermediates are identified (Dobson *et al.*, 1994). The earliest detectable species is a partially collapsed fluctuating state with substantial secondary structure but with few stable tertiary interactions. The second type of intermediate is populated in the next stage of folding in which persistent structure evolves. This structure has amide protection from exchange in only the α-domain. The β-domain is still unstable and rapidly fluctuating (5–10 ms). By about 200 ms into folding process, a single major intermediate remains in which the α-domain has formed a highly stable structure with a hydrophobic well-developed core. The native conformation of the enzyme forms in the slowest step with a time range of about 350 ms. In the majority of molecules, this occurs concomitantly with the formation of the stable structure in the β-domain.

13.6.3.2 Folding mechanism.

Three folding mechanisms have been proposed (Daggett and Fersht, 2003):

1. *Framework model*: The protein folding begins with the secondary structures. This is followed by docking of the pre-formed secondary structure units to produce the native, folded macromolecule (Kim and Baldwin, 1990). For small proteins with stable secondary structure(s), they tend to adopt α-helical and turn or β-hairpin structures. These structures may start the folding process.
2. *Hydrophobic collapse model*: The formation of hydrophobic clusters drives compaction of the protein so that the polypeptide chain folds in a confined volume narrowing the conformational search to the native state (Kim and Baldwin, 1990). The expulsion of water from the burial of nonpolar surfaces provides the hydrophobic driving force and the secondary structures are formed during hydrophobic collapse.
3. *Nucleation-condensation model*: This model integrates both the framework and hydrophobic collapse mechanisms in that the folding transition state involves the formation of the long-range and other native hydrophobic interactions, which stabilize the weak secondary structure (Fersht, 1997). The transition state probably resembles a distorted form of the native structure, with the least distorted part being loosely defined as the nucleus and the distortion tending to increase with increasing distance from the nucleus (Daggett, 2002).

Computer simulation studies, which incorporate the results of mutational analysis of folding kinetics, have provided a molecular description of the conformations that are rate-limiting in the folding of a given protein known as transition state ensemble (TSE). Examination of TSEs has shown that establishing the correct overall topology of the polypeptide chain is a crucial aspect of protein folding (Lindorff-Larsen *et al.*, 2004). Despite the many different ways in which a given topology can be generated, individual protein uses interactions between a specific and limited set of residues to define the fold of TSEs with overall topologies that are similar to those of their corresponding native states. Although the TSE structures have the key features of the protein architecture that define the overall native fold, the degree of structural similarity to the native state of typical members of the TSE is low. The formation of a specific network of interactions between amino acid residues that can be described as an extended nucleus is often found within the hydrophobic core, and thus restricts the region of conformational space available to the protein chain for the definition of the TSE topology. Once the overall chain topology has been established, any local secondary structure not present in the TSE can readily form during the last stages of the folding reaction in which the side chains lock together to generate the closely packed native state. For larger proteins the topology can develop locally within domains of the structure that subsequently dock together, with a few key residues enabling the correct overall architecture to be established. The amino acid sequences of cellular proteins seem to have been selected during evolution to have properties such as the ability to fold efficiently to the closely packed structures that are typical of the majority of these proteins (Lindorff-Larsen *et al.*, 2005).

13.6.4 Molecular chaperone in cytosolic protein folding

The final folded conformation of a protein *in vivo*, as *in vitro* is determined by its amino acid sequence. Proteins begin to fold as they are being synthesized on ribosomes. Nevertheless, proteins may be assisted in folding by a family of helper proteins known as molecular chaperones (Frydman, 2001; Hendrick and Hartl, 1993; Sigler *et al.*, 1998). Molecular chaperones comprise several structurally unrelated protein families, and many of them are also known as heat-shock proteins (HSP, induced by a variety of cellular stresses). Molecular chaperones bind to and stabilize the non-native conformations of proteins then facilitate their correct folding by releasing them in a controlled environment. Molecular chaperones are characterized by the ability to recognized structural elements exposed in unfolded or partially denatured proteins (e.g. hydrophobic surfaces) without an apparent sequence preference. The role of molecular chaperones in cooperation with co-chaperones, is prevention of the incorrect intermolecular association of unfolded polypeptide chains, maintenance of these polypeptide chains in the folding competent state and provision of a favorable environment for the correct protein folding.

Two molecular chaperone systems have been extensively studied (Sigler *et al.*, 1998). Hsp70 and its homologs recognize short extended peptide segments enriched in hydrophobic amino acids and therefore interact with most unfolded polypeptides. The ATP-dependent binding and release of unfolded polypeptide from Hsp70 is modulated by various co-chaperones. By contrast, the ring shaped (double toroids) chaperonins provide a central cavity in which the ATP-dependent folding of a single polypeptide chain takes place within the hydrophilic folding cage. Upon binding of ATP and, in some cases, a co-chaperonin, the chaperonin is converted to an active device that may further unfold a polypeptide chain and then release it into a shield environment (central chamber with expanded volume and hydrophilic) favorable for folding the polypeptide chain into the native state. These two folding systems in *E. coli* are represented in Figure 13.21.

13.7 BIOENGINEERING OF BIOMACROMOLECULES

13.7.1 Recombinant DNA technology

13.7.1.1 DNA Cloning. The elucidation of molecular events, enzymology of genetic duplicative process (Burrell, 1993; Eun, 1996) and knowledge of dissecting nucleotide sequences in DNA (Ansorge *et al.*, 1997) provide impetus for the development of recombinant DNA technology (Greene and Rao, 1998; Watson *et al.*, 1992). The main objective of recombinant DNA technology or molecular cloning is to insert a DNA segment (gene) of interest into an autonomously replicating DNA molecule, known as the cloning vector, so that the DNA segment is replicated with the vector (Lu and Weiner, 2001; Sambrook *et al.*, 1989). The result is a selective amplification of that particular DNA segment (gene). The technology entails the following general procedures (Berger and Kimmel, 1987; Brown, 2001):

1. cutting DNA at precise locations to yield the DNA segment (gene) of interest by use of sequence specific restriction endonuclease (restriction enzyme);
2. selecting a proper vector and cutting the vector preferably with the identical restriction enzyme;

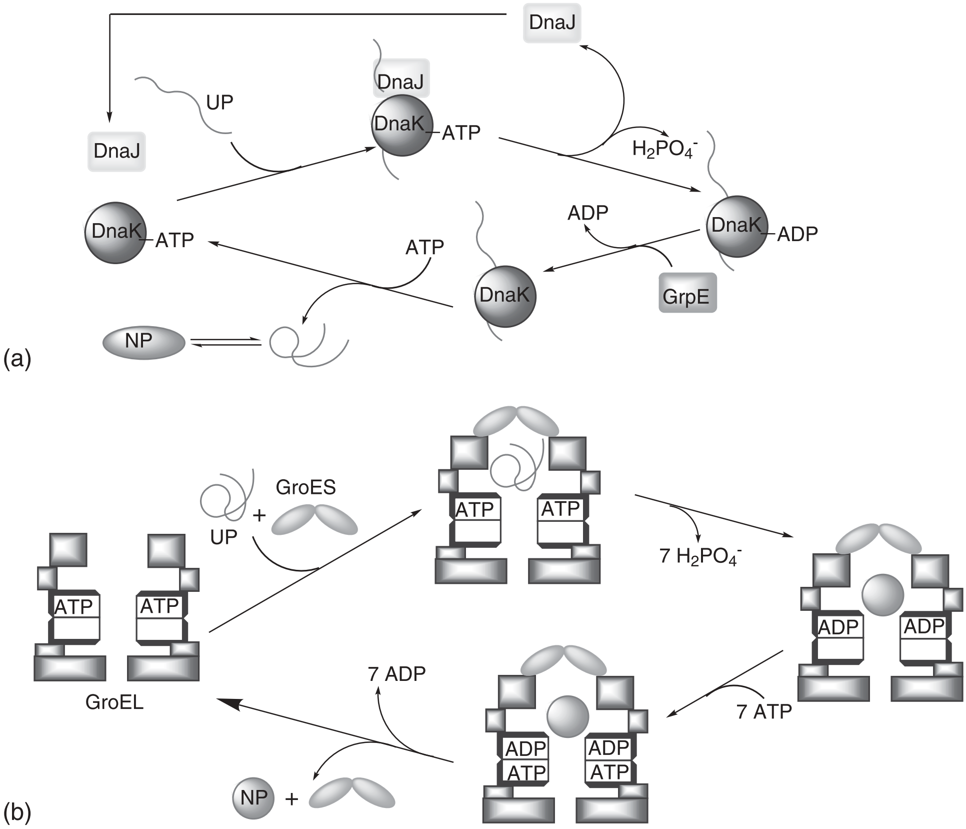
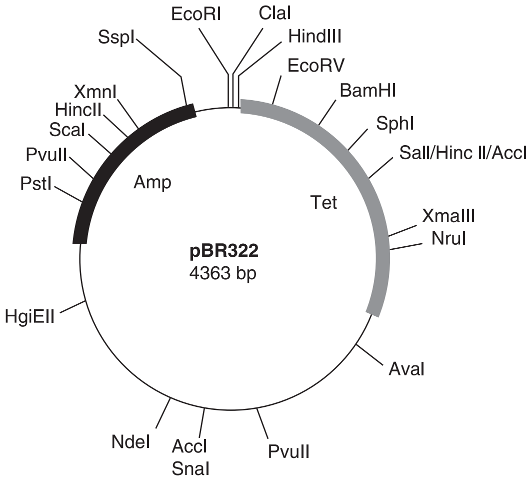


Figure 13.21 Schemes for Hsp70 and chaperonin systems in protein folding. The proposed schemes for *E. coli* Hsp70 (DnaK) and chaperonin (GroEL) actions in protein folding are depicted. In the DnaK action (a), the binding of the unfolded polypeptide chain (U) to DnaK-DnaJ (DnaJ is a co-chaperone which can binds U as well) protects U from aggregation. Whenever U is released, folding to the native protein (NP) occurs. In the GroEL-mediated folding (b), the co-chaperonin (GroES) binds as a cap at one end of the double-toroid shaped chaperonin (GroEL) to form GroEL(ATP)$_7$-GroES chaperonin complex which provides the shielded environment (central chamber/cavity) for the ATP-dependent protein folding. GroEL subunits (seven for GroEL while eukaryotic cytosolic chaperonin rings contain eight) fold into three domains, an ordered highly α-helical equatorial domain, an apical domain that surrounds the opening at the ends of the central channel and a small slender intermediate domain linking the equatorial and the apical domains. The heptameric GroES, each consisting of a core β-barrel with two β-hairpin loops arches upward and inward to form the top of the dome. The bullet-like GroEL-GroES structure has a cavity ($\sim 85 \times 10^3$ Å^3), just large enough for a native protein with a molecular size of 70 kDa. The chaperonin folding cycle consists of (1) binding of polypeptide chain by GroEL, (2) release of polypeptide chain into the central chamber and initiation of folding accompanied by binding of GroES and ATP (*cis*-active state), (3) hydrolysis of ATP, priming the *cis* assembly to release bound peptide, (4) binding of ATP to the *trans* ring and trigger the discharge of GroES and folded protein (NP)

3. construction of the recombinant DNA by inserting the DNA segment of interest into the vector by joining the two DNA fragments with DNA ligase;
4. introduction of the recombinant DNA into a host cell that can provide the enzymatic machinery for DNA replication;

5. selecting and identifying those host cells that contain recombinant DNA;
6. identification of the recombinant DNA.

Various bacterial plasmids, bacteriophages and yeast artificial chromosomes are used as cloning vectors (Brown, 2001), such as pBR322.



VectorDB (http://genome-www2.stanford.edu/vectordb/) offers characterization and classification of nucleic acid vectors. A catalog of vectors is available from American Type Culture Collection (ATCC) at http://www.atcc.org/ and PlasMapper (http://wishart.biology.ualberta.ca/PlasMapper/) is a web server for drawing and annotating plasmid maps. At the heart of the general approach to generating and propagating a recombinant DNA molecule is a set of enzymes (subsection 13.3.2), which synthesize, modify, cut and join DNA molecules.

The average size of the DNA fragments produced with a restriction enzyme depends on the frequency with which a particular restriction site occurs in the original DNA molecule and on the size of the recognition sequence of the enzyme. In a DNA molecule with a random sequence, the restriction enzyme with a six-base recognition sequence produces larger fragments than the enzyme with four-base recognition sequence. Another important consideration is the preservation of the original reading frames in the recombinant DNA.

13.7.1.2 Polymerase chain reaction. Polymerase chain reactions, which apply the DNA polymerase replicating system to synthesize DNA, offers a convenient method of amplifying the amount of a DNA segment (Innis *et al.*, 1999; Mullis *et al.*, 1994; Newton and Graham, 1997).

Basic principles of the polymerase chain reaction: The polymerase chain reaction (PCR) procedures have extended the range of *in vitro* DNA amplification by up to about 20 kb. For longer targets, *in vivo* cloning methods must be used. The simplest amplification scheme for *in vitro* DNA amplification is successive cycles of priming, chain extension and product denaturation. If these steps are carried out efficiently, the result is a linear increase in the amount of product strand with increasing cycles of amplification. This scheme, called linear amplification, is useful in preparing DNA samples for DNA sequencing. The step up in complexity from linear amplification is the use of two antiparallel primers. The primer selection for PCR can be accessed at Pimer3 (http://genome.wi.mit.edu/cgi-bin/primer/primer3wwwcgi), PROBEmer (http://probemer.

cs.loyola.edu/) or Web Primer (http://genome-www2.stanford.edu/egi-bin/SGD/web-primer). MPDB (http://www.biotech.ist.unige.it/interlab/mpdb.html) compiles synthetic oligonucleotides useful as primers or probes. The target DNA is denatured and two antiparallel primers are added. These must be in sufficient excess over target that renaturation of the original duplex is improbable, and essentially all products are primers annealed to single-stranded templates. Hence it doubles the number of targets present in the starting mixture. Each successive round of DNA denaturation, primer binding and chain extension will, in principle, produce a further doubling of the number of target molecules. Hence the amount of amplified product grows exponentially as 2^n, where n is the number of amplification cycles. In practice, the actual PCR efficiencies typically achieved are not perfect but they are remarkably good. We can define the efficiency, E, for n cycles of amplification by the ratio of product, P, to starting material, S as

$$P = S(1 + E)^n$$

Actual efficiencies turn out to be in the range of 0.60 to 0.95.

The most widely used DNA polymerase in PCR is from *Thermus acquaticus*, the Taq polymerase. This enzyme has an optimal temperature for polymerization in the range of 70 to 75°C. It extends DNA chains at a rate of about 2 kb per minute. It is fairly resistant to the continual cycles of heating and cooling required for PCR. The half-life for thermal denaturation of Taq polymerase is 1.6 h at 95°C. When very high denaturation temperatures are needed, as in the PCR amplification of very G+C-rich DNA, more thermal-stable polymerase such as the enzyme from *Thermococcus litoralis* with a half-life of 1.8 h at 100°C or the enzyme from *Pyrococcus furiosis* with a half-life of 8 h at 100°C, can be employed. However, these enzymes are not as processive as Taq polymerase, which makes it more difficult to amplify long templates.

With Taq polymerase typical PCR cycle parameters are:

DNA denaturation	92–96°C for 30 to 60 seconds
Primer annealing	55–60°C for 30 seconds
Chain extension	72°C for 60 seconds

The protocol has been automated through the introduction of thermal cyclers that alternately heat the reaction mixture to dissociate the DNA, followed by cooling, annealing of primers and another round of DNA synthesis. The number of cycles used depends on the particular situation and sample concentrations, but typical values when PCR is efficient are 25 to 30 cycles. Thus the whole process takes about an hour. Typical sample and reagent concentrations used are:

Target DNA	$<10^{-15}$ mol
Primer oligonucleotide	2×10^{-11} mol
Dexoynucleoside triphosphates	2×10^{-8} mol each

These concentrations imply that the reaction will eventually saturate once the primer and triphosphates are depleted.

Taq polymerase has no 3′-proofreading exonuclease activity. Thus it can misincorporate bases. However, the enzyme does have a 5′-exonuclease activity and therefore will nick translate. The most serious problem generally encountered is that Taq polymerase can add an extra nontemplate-coded A to the 3′-end of DNA chains. It can also use a related activity, making a primer dimeric that may or may not contain additional uncoded residues. Once such dimeric primers are created, they are efficient substrates for further amplification. These primer dimers are a major source of artifacts in PCR. However, they are usually short and can be removed by gel filtration or other sizing methods to prevent their interference with subsequent uses of the PCR reaction product.

Some PCR variations: A number of variations on the basic PCR reaction increase the utility of the technique. One of these is a convenient approach to the simultaneous analysis of a number of different genomic targets called multiplex PCR. Each target to be analyzed is flanked by two specific primers. One is ordinary but the other is tagged with a unique 20-bp overhanging DNA sequence. PCR can be carried out separated or together, and then the resulting products pooled. The tagged PCR products are next hybridized to a filter consisting of a set of spots, each of which contains the immobilized complementary sequence of one of the tags. The unique 20-bp duplex formed by each primer sequence will ensure that the corresponding PCR products become localized on the filter at a predetermined site. Thus the overall results of the multiplex PCR analysis will be viewed as positive or negative signals at specific locations on the filters. This approach, where amplified products are directed to a known site on a membrane for analysis, dramatically simplifies the interpretation of the results and makes it far easier to automate the whole process.

For DNA sequencing, it is highly desirable to produce a single-stranded DNA product. To produce single strands in PCR, a simple approach called asymmetric PCR can be used. Here ordinary PCR is carried out for a few less cycles, say 20. Then one primer is depleted or eliminated, and the other is allowed to continue through an additional 10 cycles of linear PCR. The result is a product that is almost entirely single stranded. Clearly we can use whichever strand we want by appropriate choice of primer.

Another PCR variant is called DNA shuffling (Stemmer, 1994). Here the goal is to enhance the properties of a target gene product by *in vitro* recombination. Suppose that a series of mutant genes exist with different properties and the goal is to combine them in an optimal way. The genes are randomly cleaved into fragments, pooled, and the resulting mixture is subjected to PCR amplification using primers flanking the gene. Random assembly of overlapping fragments will lead to products that can be chain extended until full length reassembled genes are produced. These then support exponential PCR amplification. The resulting populations of mutants are cloned and characterized by some kind of screen or selection in order to concentrate those with the desirable properties. An interesting alternative method to shuffling DNA segments uses catalytic RNAs (Mikheeva and Jarrell, 1996).

The amplification of cDNA by combining a reverse transcriptase (RT) reaction with PCR is known as reverse transcription polymerase chain reaction (RT-PCR). A first-strand of cDNA is generated from RNA with RT in the presence of dNTP, Mg^{2+} by the use of either gene-specific RT primers or random hexamer primer at 50°C for 1 h. The reaction mixture is heated (e.g. 70°C for 15 min) and cooled. The resulting gene-specific RT reaction mixture is column purified and subjected to PCR amplification of cDNA.

The PCR amplification technique is also used for signal generation in antibody-based immunoassays. *Immuno-PCR* (IPCR) is based on chimeric conjugates of specific anitbodies and DNA fragment, which is used as marker to be amplified by PCR for signal generation (Niemeyer *et al.*, 2005). In the two-sandwich approach, *bis*-biotinylated dsDNA and streptavidin (STV) are assembled as the cross-linking agent, which provides binding sites for biotinylated antibodies. An affinity interaction between antigens and antibodies retains the antibody-DNA conjugates comprising DNA marker, which is, in turn, amplified by PCR for signal generation.

Total genome amplification method: One approach to PCR sampling of an entire genome is the method of primer extension preamplification (PEP). A mixture of all possible 4^n primers of length n is generated by automated oligonucleotide synthesis, using at each step all four nucleotides rather than just a single one. This extremely complex mixture is then used as a primer. Although the concentration of any one primer is exceedingly small, there are always enough primers present that any particular DNA segment has a reasonable chance of amplification. A reasonable success was reported for using this approach with n = 15 to amplify at least 78% of the genome by using 50 cycles of amplification (Arnheim and Erlich, 1992).

Real time PCR: Sometimes referred to as kinetic PCR, real time PCR is a process by which a fluorescent signal is generated as the target sequence is amplified, and this fluorescent signal is measured at each cycle of amplification during the PCR process (Heid, 1996; Orlando *et al.*, 1998). Two fundamental techniques are commonly used to generate the fluorescent signal:

1. *The use of fluorescence dye such as SYBR Green which binds dsDNA*: Any amplicon that is synthesized will bind the dye and generate a fluorescent signal that is detected.
2. *The use of a sequence specific fluorescent probe*: The fluorescent probe (FP) hybridizes to the target sequence after the denaturation step. The probe uses fluorescent resonance energy transfer technology, whereby the emission spectrum of a fluorescent reporter dye at its 5′ end is effectively quenched by a second fluorescent dye at its 3′ end when the probe is intact. The fluorescent signal is produced during primer extension when the 5′ end is displaced by the amplicon and being cleaved by the nick specific 5′ nuclease.

Real time RT-PCR has significantly simplified the process of producing reproducible quantitation of low-abundance mRNA and transcriptional profiling (Rajeevan *et al.*, 2001). Real time quantitation of mRNA by RT-PCR is defined by C_T (threefold cycle number) at a fixed threshold where PCR amplification is still in the exponential phase and the reaction components are not rate-limiting. Standard curves are constructed for each amplicon from which the precise copy number of mRNA transcripts is calculated or for selected ones from which the unknown sample is estimated by normalizing to the input amount of a reference gene (Aberham *et al.*, 2001).

13.7.1.3 Site-directed mutagenesis. Most methods for specifically mutating the codon for a specific amino acid in a protein are based on the procedure of oligodeoxyribonucleotide/site-directed mutagenesis (Smith, 1985; McPherson, 1991). The gene is cloned into a vector and one of the constituent circles of ssDNA is isolated. A short oligodeoxyribonucleotide complementary to the region of the gene to be mutated (with a single-base or double-base mismatch designed to change the codon for the target amino acid residue into the codon for the desired mutant residue) is synthesized. This oligodeoxyribonucleotide is annealed to the gene in the single-stranded vector and used as a primer for polymerase reaction (the Klenow fragment). The replicated strand is ligated to yield a hetereoduplex containing one strand of mutant and one strand of wild-type DNA. The hetereoduplex is used to transform a host and produce colonies of cells that each contains either the vector with the mutant or the vector with the wild-type gene. The colonies are screened for the desired mutant.

In a more efficient PCR method, a circular plasmid, and two mutant primers that are complementary in sequence at the site of mutation is used to perform PCR amplification. In the first cycle of mutation, each primer introduces the mutation into two respective strands and the mutants are subsequently amplified in the following rounds of amplification because all primers are complementary to the mutants. All mutant progeny can be obtained after 20 cycles of PCR.

13.7.2 RNA Engineering

13.7.2.1 Antisense RNA. Antisense RNA is defined as a short RNA transcript that lacks coding capacity, but has a high degree of complementarity to another RNA or DNA, which enables the two to hybridize. Natural antisense transcript (NAT) is a simple RNA containing sequence that is complementary to other endogenous RNA. It can be transcribed in *cis* from the opposing DNA strand at the same genomic locus (*cis*-NAT) or in *trans* from separate loci (*trans*-NAT). The consequence is that such antisense, or complementary RNA can act as a repressor of the normal function or expression of the targeted RNA (Yelin *et al.*, 2003). Such species have been detected in prokaryotic and eukaryotic cells with suggested functions concerning RNA-primed replication of plasmid DNA, transcription of bacterial genes and messenger translation in eukaryotes and prokaryotes. Since ribosomes cannot translate dsRNA, the translation of a given mRNA can be inhibited by a segment of its complementary sequence, so-called antisense RNA. A short dsRNA with siRNA activity can be synthesized and transfected into the cell. After incorporating into the RISC, the antisense strand of siRNA performs gene silencing (RNAi) by degradating mRNA with segment of sequence complementary to the sequence of the antisense RNA strand (McManus and Sharp, 2002). siRNA can be prepared for gene silencing by:

a) chemical synthesis;
b) *in vivo* transcription;
c) digestion of long dsRNA by an RNase III (Dicer family);
d) expression in cells from an siRNA expression vector; and
e) expression in cells from a PCR derived siRNA expression cassette.

AOBase (http://bio-inf.net/aobase) is an online facility for antisense oligonucleotide selection and design, and siRNAdb (http://sirna.cgb.ki.se/) offers a search engine for siRNA.

For example, fruit ripening is controlled by the plant hormone, ethylene (C_2H_4), the product of a metabolic pathway whose rate-determining step is catalyzed by 1-aminocyclopropane-1-carboxylate synthase (ACC synthase). The constitutive expression of the antisense RNA of the ACC synthase gene in fruit would therefore be expected to inhibit fruit ripening. In fact the transgenic expression of this antisense RNA in tomato plants prevents tomato ripening, an effect that can be reversed by the administration of ethylene. This effect raises the prospect that fruits and vegetables can be ripened on demand thereby preventing spoilage during storage and transportation.

13.7.2.2 Engineering ribozyme. Ribozymes (section 11.7) can now be engineered into a multiple turnover RNA enzymes (Zang and Cech, 1986; Pyle and Green, 1994; Ekland *et al.*, 1995). The naturally occurring self-cleaving RNAs are single stranded RNA molecules, but can be made into true enzymes exhibiting multiple substrate turnover simply by division into two strands of RNA. Most ribozymes facilitate intramolecular reactions, self-splicing or self-cleavage, so a physiological concentration of substrate cannot be defined. Tight binding and slow release of intermediates may play a biological role in self-splicing and this could explain why ribozymes engineered to act with multiple turnover are typically easily saturated and have overall turnover limited by slow product release (Herschlag and Cech, 1990). When substrate (or product) recognition involves base pairing, turnover can often be improved simply by introducing mismatches or shortening the recognition strand of ribozyme (Zang *et al.*, 1994; Fedor and Uhlenbeck, 1992). When substrate recognition involves tertiary interactions, mutations can improve turnover by weakening these interactions (Young *et al.*, 1991).

13.7.3 Protein engineering

Protein engineering is a hybrid discipline wherein recombinant DNA technology, conventional protein chemistry and a host of biochemical, chemical and physical techniques are applied to design, produce and investigate proteins with two purposes.

1. dissection of the structure and activity of existing proteins by making systematic alternations of their structures and examining the changes in their function and properties; and
2. production of novel proteins *de novo* or altered proteins to give useful changes in activities and/or properties.

Two classes of functional proteins will be considered. This subsection considers engineering enzymes and the next subsection deals with engineering antibodies.

13.7.3.1 Engineering enzyme by site-directed mutagenesis. Site-directed mutagenesis has been applied to investigate enzyme catalytic mechanism, substrate/ligand binding, protein folding and subunit interactions. Mutant enzymes aimed at investigating structure-function relationships should minimize reorganization of the structure of the enzymes, either locally or globally, for which:

- Choose a mutation that decreases the size of a side chain or leads to an isosteric change.
- Choose deletion mutation over addition mutation.
- Avoid creating buried unpaired charges.
- Minimize an alternation in which involves multiple interactions.
- Avoid adding new functional groups to side chains.

The ideal mutation for enzyme studies is a nondisruptive deletion, i.e. only removing an interaction without causing a disruption or reorganization of protein structure. The alternations (mutations) causing least structural effect are:

Mutation	Change	Probe
Ile→Val, Ala→Gly, Thr→Ser	Shortening chain by CH_2	Hydrophobic interaction
Val→Ala, Leu→Ala, Ile→Ala	Shortening/debranching	Hydrophobic/steric effect
Ser→Ala, Tyr→Phe, Cys→Ala	Deletion of OH/SH	Hydrogen bonding
Asp→Asn, Glu→Gln	Isosteric, COO^- to $CONH_2$	Ionic/hydrogen bond interactions
His→Asn, His→Gln	Deletion of imidazole ring	N^{δ}(N)/N^{ε}(Q) of imidazole
Cys→Ser or Ser→Cys	Isoelectronic, SH to OH	Steric/dipole effect

The application of site-directed mutant enzymes to investigate catalytic reaction mechanisms has been considered (subsection 11.4.2). The site-directed mutagenesis has been used to fine-tune enzyme activity, to redesign enzyme specificity/catalysis and to improve enzyme properties such as enhancement of enzyme stability (Table 13.11).

TABLE 13.11 Examples of some effects of site-directed mutagenesis on enzymes

Enzyme	Mutant	Observation	Effect: possible explanation
Trypsin, rat	G216A G226A GG216-226AA	Relative (Arg/Lys) k_{cat} & K_m = 1.0/0.3 and 30/30 for G216A; 0.01/0.1 and 40/25 for G226A; 0.001/0.005 and 15/2 for double mutant.	Substrate specificity: G216A mutant operates more selectively on Arg substrate and G226A mutant operates more selectively on Lys substrate. The double mutant has a very low activity.
Lactate dehydrogenase, *B. stearothermophilus*	AA236-237GG (Dm) QKP102-105MVS (Tm) QKP102-105MVS/AA236-237GG (Qm)	k_{cat}/K_m for RCOCOOH where R = CH_3, C_2H_5, C_3H_7, C_4H_9, and $(CH_3)_2CH$ are 4.2×10^6, 1.7×10^5, 6.4×10^3, 8.5×10^3, 12 for wild; 4.2×10^4, 3.2×10^4, 7.0×10^3, 4.3×10^4, 50 for Dm; 4.1×10^5, 1.8×10^5, 9.5×10^3, 8.7×10^3, 67 for Tm; 8.0×10^3, 6.2×10^4, 4.2×10^4, 2.6×10^4, 570 for Qm.	Substrate specificity: The replacement of the jaw region (substrate side-chain recognition) $^{236-237}$AA with GG enlarges the pocket. The replacement of coenzyme loop (closes over the substrate) $^{102-104}$QK with MV increases the hydrophobicity of the region and ^{105}P with S provides more flexibility to the loop so that larger substrate may access to the active site. The quintuple mutant produces broader substrate specificity.
Lysozyme, human	Y63F/W/L/A	Rms differences in side-chain atoms of active site = 0.128, 0.110, 0.148, 0.137; ΔG_R (kJ/mol) = 1.67, 3.35, 10.88, 10.46 for F, W, L, A.	Binding specificity: X-Ray study and kinetic analysis suggest the direct contact between the planar side-chain of ^{63}Y and the sugar residue at subsite B is a major determinant of binding specificity.
B. subtilis aspartate transcarbamylase	R99A	n_H (Asp) = 1.0 for wild and 1.5 for R99A	Regulatory property: Conversion of a non-cooperative interaction to an enzyme with homotropic substrate cooperativity.
Triosephosphate isomerase, yeast	N78D N14T/N78I	No appreciable lowering in k_{cat} for both mutants but the mutations compromise the stability of the enzyme to an elevated temperature, urea, alkaline pH and proteases.	Subunit interaction: The replacement of Asn-78 at the subunit interface does not appreciably affecting the catalytic efficiency, but the mutants become less resistant to denaturation and inactivation.

TABLE 13.11 ***continued***

Enzyme	Mutant	Observation	Effect: possible explanation
Phospholipase A_2, pancreas	Y73S D99N	Disappearance/broadening of nonexchangeable NH protons (δ7.8–9.5 ppm region) in Y73S/D99N by NMR but without substantial difference in CD spectra between wild and mutants.	Conformation and folding: Tyr-73 and Asp-99 are part of the conserved catalytic network at the active site. Site-directed mutations of these two residues result in highly flexible conformations characteristic of the molten globule which still preserves secondary structures.
Glyceraldehyde-3-phosphate dehydrogenase, chicken	G215A G316A	K_{cat} (s^{-1}), T_{50} (°C) and $[urea]_{50}$ (M) are 231, 50 and 1.4 for wild; 231, 47 and 1.2 for G215A; 233, 57 and 1.8 for G316A.	Stability: Single mutation of Gly (in helical regions) to Ala does not affect catalytic activity but destabilizes or destabilizes the enzyme. Gly-316 is located adjacent to a rather large internal cavity and substitution with Ala may improve the packing and stability.
Xylose isomerase, *Actinoplanes missouriensis*	K253R K253Q K309R K319R K323R K309R/K319R/K323R	Half-life ($t_{1/2}$) of wild at 84°C is 2.7–3.0 hr. $t_{1/2}$ of K253R and K253Q are 3.7 hr and less than 1 min. The triple mutant (with $t_{1/2}$ of 7 hr) is more stable than either of single mutants.	Thermal stability: The K→R mutations do not affect the enzymatic activity appreciably. The single mutant displays modest improvement in thermal stability with a greatly enhanced stability for the triple mutant. Electrostatic interactions on the protein surface may provide stabilizing contribution.

Notes: 1. Mutants are expressed (one-letter abbreviation for amino acids) as: wild-#position-mutant, i.e. G216A denotes the replacement of Gly216 for Ala in the single mutant; QKP102-105MVS for triple mutant (Tm) in which $^{102-105}$QKP are replaced by MVS and QKP102-105MVS/AA236-237GG represents quintuple mutant (Qm) in which $^{102-105}$QKP are replaced by MVS and $^{236-237}$AA by GG. Multiple single mutants, Y63F, Y63W, Y63L and Y63QA are represented by Y63F/W/L/A.

2. Data taken form: Craik *et al*. (1987) for rat trypsin; Wilks *et al*. (1990) for *B. stereothermophilus* lactate dehydrogenases; Muraki *et al*. (1992) for human lysozyme in which the active site residues refer to E35, D53, W64, D102 and W109 and $\Delta G_R = -RTIn[(k_{cat}/K_m)_{mutant}/(k_{cat}/K_m)_{wild}$; Stebbins and Kantrowitz (1992) for *B. subtilis* aspartate transcarbamylase; Casal *et al.* (1987) for yeast triosephosphate isomerase; Dupureur *et al*. (1992) for pancreatic phospholipase; Mrabet *et al.* (1992) for *Actinoplanes missouriensis* xylose isomerase.

Active sites of enzymes and binding sites of proteins are a general source of instability because they contain groups that are exposed to solvent in order to bind substrates and ligands. Thus there seems a trade-off between stability and activity. An approach is to find mutagenic amino acids that can be altered to enhance protein stability. For example, an improvement in the stability of the two largest volume industrial enzymes, glucose isomerase and α-amylase, would greatly enhance their performance in the process for conversion of starch to high-fructose corn syrup. A general strategy to improve the enzyme stability cannot be formulated because it depends on the individuality of the protein and the environments of targeted amino acid residues. The ability to design proteins with enhanced thermal stability is based on the aggregate of structural information, which may be derived from the comparative studies of thermophilic proteins with their mesophilic counterparts. The thermal stability of many thermophilic proteins seems to correlate with their ability to strengthen the nonpolar character of the internal nucleus of the proteins (Matthews, 1995). In large multisubunit proteins, the large interior surface may provide opportunity for improvement in packing and stability.

13.7.3.2 Engineering enzyme: Redesigning enzyme. Redesigning enzyme concerns with the conversion of an enzyme from its existing characteristics to new and different catalytic characteristics. DNA recombinant technology offers unprecedented opportunity for redesigning enzymes with desired specificity and activity. However, knowledge/information on sequences and 3D structures of the sample and target enzymes (or their homologues) are essential for identifying which individuals or clusters of amino acid residues are to be mutated (deleted, inserted and/or substituted). Some approaches for redesigning enzymes are shown in Table 13.12.

TABLE 13.12 Examples of redesigning enzymes

Enzyme	Mutant	Observation	Design: possible explanation
Lipoamide dehydrogenase, *E. coli*	A: EMPD203-206VRKH/P210R B: G85A/G89A/EMPD203-206VRKH/P210R	k_{cat}/K_m ($s^{-1}M^{-1}$) for NAD^+ *vs* $NADP^+$ = 1.25×10^6 *vs* nil for wild; 4.32×10^4 *vs* 5.82×10^6 for A; 6.32×10^3 *vs* 1.58×10^6 for B.	Introduction of $NADP^+$ activity: Systematic replacement of amino acids in the βαβ-fold of the NAD^+-binding domain which is unfavorable to NAD^+ but favorable to $NADP^+$, thus converts the $NADP^+$ inactive wild-type to mutants with high $NADP^+$ activity.
*Iso*Citrate dehydrogenase, *thermos thermophilus*	A: KY283-284DI/NVI288-290GIA B: R231A/KY283-284DI/NVI288-290GIA	k_{cat}/K_m ($s^{-1}\mu M^{-1}$) for NAD^+ *vs* $NADP^+$) = 9.9 vs 0.013 for wild; 0.018 *vs* 0.19 for A; 0.013 vs 0.88 for B.	Interconversion of nicotinamide coenzyme specificity: Module replacement of the loop sequences in the βαβ-fold of ICDH to the homologous NAD^+-enzyme convert the $NADP^+$-dependent ICDH to NAD^+-dependent ICDH.
Lactate dehydrogenase, *B. stereothermophilus*	Wild D197N Q102R	k_{cat} (s^{-1}) and K_m for Pyr/OAA = 250/6.0 and 0.06/1.5 for wild; 90/0.50 and 0.66/0.15 for D197N; 0.9/250 and 1.8/0.06 for Q102R.	Conversion of lactate dehydrogenase to malate dehydrogenase: The replacements of D197 and Q102 at the binding site change the substrate specificity of lactate DH from pyruvate to oxaloacetate.

Notes: Data taken from: Yaoi *et al*. (1996) for *E. coli* lipoamide dehydrogenase; Yaoi *et al.* (1996) for *Thermus thermophilus iso*citrate dehydrogenase (ICDH); Wilks *et al*. (1988) for *B. stearothermophilus* lactate dehydrogenase.

Conversion of coenzyme specificity. The conversion of nicotinamide coenzyme specificity of oxidoreductases is one of the most frequently performed operations because amino acid sequences and 3D structures of many dehydrogenases are known. The two coenzymes (NAD^+ and $NADP^+$) differ only by the presence of a phosphate group esterified to the 2′-hydroxyl group of the adenylyl moiety of $NADP^+$. Analysis of coenzyme binding in dehydrogenases shows a structural homology within the βαβ-fold fingerprint region, which contains amino acid sequences that are characteristic for NAD(H) or NADP(H) binding. Prominent in the coenzyme binding is the presence of two amino acid residues, Asp or Glu for NAD(H)-preferred dehydrogenases and Arg for NADP(H)-dependent enzymes. These structural information provide useful guide for redesigning enzymes with converted coenzyme specificity.

Conversion of substrate specificity. No general strategy can be formulated for converting enzyme from catalyzing one set of substrates to another set of substrates because it depends on the different elements of substrate molecules that interact with specific amino acid residues of the enzyme. For example, the conversion of lactate dehydrogenase to malate dehydrogenase is achieved by mutating the amino acid residues (Gln102 and Thr246), which interact with the methyl group of pyruvate to the residues (Arg and Gly), which favor CH_2COO^- functionality. The mutant T246G is aimed at enlarging the binding pocket for the bulkier carboxymethyl group and the mutant Q102R introduces a positive charge, which may interact with the γ-carboxylate anion of oxaloacetate. The conversion of lactate dehydrogenase to malate dehydrogenase is indicated by greatly enhanced affinity and catalytic efficiency of the mutant for oxaloacetate.

Introduction of activator ion/prosthetic group. Metal sites can be engineered into naturally occurring folds. New sites forming a given ligation geometry are found by searching existing structures for backbone geometries consistent within the conformational needs of the ligating groups.

Shift in the equilibrium of catalytic reaction (subtiloligase, glycosynthetase). The protease, subtilisin is converted into subtiloligase, which catalyzes polypeptide synthesis (Jackson *et al.*, 1994). The key active-site residues involved in glycosidase catalyses are a pair of carboxylic acids. By manipulating the locations or even the presence of these carboxyl side chains in the active site, this has allowed the development of glycosynthetases, mutant glycosidases that are capable of synthesizing oligosaccharides but unable to degrade them (Ly and Withers, 1999). Such mutants are useful in the enzymatic synthesis of oligosaccharides.

Shift the reaction path (cyclophilin). Cyclophilin, which catalyzes the *cis-trans* isomerization of peptidyl-proline bonds, is converted into protease by an engineered introduction of the proteolytic triad, Asp-His-Ser (Quéméneur *et al.*, 1998).

For the production purpose, abolishment of feedback control and fusion of sequential enzymes may be desirable. The recombinant technology is the method of choice when the redesigning involves substitutions between coded amino acids. However, for substitutions involving artificial, noncoded amino acids or their analogues/derivatives (e.g. post-translationally modified amino acids, isotopic label of specific residue), chemical methods in particular semi-chemical synthesis becomes necessary (Chaiken, 1981). The fragments/chains of modified, synthetic or expressed polypeptides are enzymatically or chemically ligated (Muir, 2003).

13.7.4 Antibody engineering

13.7.4.1 Assembly of immunoglobulin genes. Structures of immunoglobulins (Igs) have been well-studied (Alzari *et al.*, 1988; Davies and Chacko, 1993) and a wealth of their DNA sequence data is available. Some useful immunological databases (DB) are available online. They include genetic and clinical DB of human major histocompatibity complexes (dbMHC) at http://www.ncbi.nlm.nih.gov/mhc/, functional molecular immunology DB (FIMM) at http://research.i2r.a-star.edu/sg/fimm, curated DB of haptens (Haptendb) at http://imtech.res.in/raghava/haptendb/, International immunogenetics information system for Igs, T-cell receptors and MHC (IMGT) at http://imgt.cines.fr/, and IMGT associated databases. The immune system has a capacity to generate diverse antibodies with a virtually unlimited variety of binding sites against almost any antigens. This antibody diversity (Tonegawa, 1988) is generated by genetic recombination among a relatively few gene segments encoding the variable region of Ig chains via intrachromosomal recombination and by an extraordinarily high rate of Ig gene mutation (French *et al.*, 1989) during B cell differentiation. The Ig genes are highly evolved for maximizing protein diversity from a finite amount of genetic information, which is scattered among multiple genes, segments along a chromosome in germline cells. During the formation of B lymphocytes, these segments are brought together and assembled by DNA rearrangement or gene reorganization to generate a variety of protein isoforms from a limited number of genes.

1. *Assembly of an L-chain gene by combining three separate genes*: The constant region of the L-chain is encoded by C_L gene and the variable-region genes are assembled from two kinds of germline genes, V_L and J_L (J for joining). Furthermore, the two families of L-chains, κ and λ are encoded by their corresponding V and J genes. In different mature B cells, V_κ and J_κ genes have joined in different combinations, and along with the C_κ genes, form complete L_κ chains with different V_κ regions. Thus the assembly of the L-chain gene occurs by DNA rearrangements that combine three genes ($V_{\kappa,\lambda}$, $J_{\kappa,\lambda}$ and $C_{\kappa,\lambda}$) to make L polypeptide chain.

2. *Assembly of an H-chain gene by combining four separate genes*: A V_H gene, which encodes the first 98 amino acids of the variable region is joined to a D (diversity) gene, which encodes amino acids 99 to 113 of the H-chain. Each V_H gene has an accompanying L_H (leader) gene, which encodes its essential leader peptide. The V_H-D gene assemblage is joined to a J_H gene, which encodes the remaining part of the variable region. In B cells, the variable region of an H-chain gene is composed of one each of the L_H-V_H genes, a D gene and a J_H gene joined head to tail. Because the H-chain variable region is encoded in three genes (L_H-V_H, D and J_H) and the joining can occur in various combinations, the H-chains have a greater potential for diversity than the L-chain variable regions that are assembled from two genes (L_L-V_L and J_L).

3. *V-J and V-D-J joining in gene assembly*: A specific nucleotide sequence adjacent to various variable-region genes may serve as joining signals. All germline V and D genes are followed by a consensus CACAGTG heptamer separated from a consensus ACAAAAACC nonamer by a short nonconserved 23-bp spacer. Similarly, all germline D and J genes are preceded immediately by a consensus GGTTTTTGT nonamer separated from a consensus CACTGTG heptamer by a short nonconserved 12-bp spacer. The consensus elements downstream of a gene are complementary to those upstream from the gene with which it recombines. These complementary consensus sequences serve as recombination recognition signals (RRS's) and determine the site of recombination between variable-region genes.

13.7.4.2 Phage display library. Phage display technology is well suited for high thorough-put generation of antibodies for diagnostic/therapeutic applications and proteomic researches for protein detection, expression profiling and functional studies (Kay *et al.*, 1996; Rader and Barbas, 1997). The selection of antibodies by phage display relies on several factors:

1. the ability to isolate or synthesize antibody gene pools to construct large, highly diverse libraries;
2. the possibility to express functional antibody fragments in the periplasmic space of *E. coli*; and
3. the efficient coupling of expression and display of the antibody protein in the *E. coli* bacteriophage.

Functionally antibody domains can be expressed on the surface of filamentous bacteriophage (e.g. M13) as a fusion to the N-terminus of the minor phage coat protein, e.g. gIIIp (pIII). The favored format of antibody fragment is single-chain Fv (scFv), a protein composed of the variable regions heavy and light chains (V_H and V_L) connected by a flexible peptide linker or Fab fragment consisting variable and constant regions of heavy and light chains of IgG (Figure 13.22).

The DNA sequence encoding the antibody domains may be cloned into the phage vector. However, phagemid (a plasmid with both a plasmid and a phage origin of replication) vectors, which are easier to manipulate and have higher transformation efficiencies, are generally favored. For the scFv library, human lymphoid cells are the ideal source of DNA sequences encoding the V_H and V_L regions of IgG. The domains are amplified by PCR using families of primers. The V_H and V_L gene sequences are then connected via linker region by a subsequent overlap extension PCR (assembly PCR) to give a complete scFv antibody DNA sequence. V_H and V_L domains are paired randomly, and this

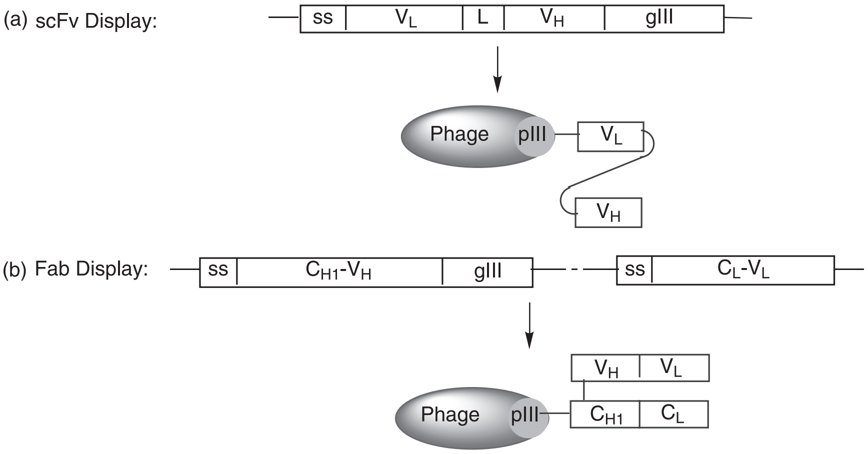


Figure 13.22 Phage display of Fv and Fab. The DNA encoding for variable regions of light and heavy chains (V_L and V_H) are linked and cloned into phasmid vector for phage display of Fv fragment (a). The bacterial signal sequence (ss) directs secretion of the expressed protein to the periplamic space and phage gene III (gIII) mediates attachment to the phage particle. The DNA encoding for the constant and variable regions of the heavy (C_{H1} and V_H) and light (C_L and V_L) chains are cloned into separate sites of the phasmid vector used for phage display. The disulfide bond connects the heavy chain attached to the phage particle and the light chain to form the Fab fragment (b)

non-native pairing can increase functional diversity in the library. The scFv sequences are amplified by PCR using primers incorporating restriction sites that facilitate cloning into the phagemid vector. After ligation into the recipient phagemid, the library is transformed into *E. coli* by electroporation. The phage library proper is obtained by rescuing the culture of the phagemid library in *E. coli* with a helper phage that will package the phagemid DNA. For Fab library construction, the V_H-C_{H1} and V_L-C_L regions are amplified and inserted into a phage display vector such that the heavy chain region is fused to the capsid gene while the light chain is inserted next to a bacterial signal sequence to direct secretion of the chain to the periplasm of *E. coli*. Within the oxidizing environment of the periplasm, a disulfide bond forms between the heavy and light chains to place the entire Fab fragment on the surface of the phage.

Specificity and affinity are desirable qualities in any antibody. The larger the library, the greater the likelihood that it will contain an antibody of high affinity having the desired specificity. Human antibody libraries readily yield clones recognizing any human protein tested. The library contains scFv with affinities better than 10^8–$10^{-10}\,M^{-1}$ for a large diversity of antigens. The selection of antibodies (against purified antigen) from phage libraries consists of two main steps: panning and screening.

During panning, an antigen is coated on to a solid surface, such as the well of a micrometer plate, and the library of phage antibodies is incubated with the target. The surface is washed, leaving behind those phages that display antibodies binding to the target antigen. Specific phages are propagated by being infected into *E. coli*. Further round(s) of panning may be applied to yield a polyclonal mixture of phage antibodies enriched for antigen-specific binders. The screening process involves subsequently converting this polyclonal mixture into monoclonal antibodies. For this process, *E. coli* cells are infected with the phage pool, plated on selective agar plates and single colonies are picked. Thus highly specific, monoclonal antibody clones are obtained, from which the antibody genes can be readily isolated for further analysis or engineering. To induce soluble scFv expression into the bacterial periplasm, clones of cells containing the pCANTAB plasmid are incubated with *i*sopropyl-β-D-thiogalactopyranoside. A portion of scFv tends to leak out of the cells into the culture medium. More scFv can generally be obtained from periplasmic extracts, released from the cells by standard procedures. The pCANTAB vector adds a hexahistidine tag to the C-terminus of the scFv, allowing easy purification by immobilized metal affinity chromatography (IMAC).

Fab and scFv fragments are attractive alternative to immunoglobulins for therapeutic applications with following advantages:

- no Fc-receptor binding activation;
- a higher penetration rate of tissues; and
- both Fab and Fv antibodies can be produced in prokaryotes.

Their rather short *in vivo* half-life can be overcome by conjugation to polyethyleneglycol molecules. The phage display format also allows facile cloning into IgG vectors, so that the human IgG from mammalian cell cultures can be expressed.

13.7.4.3 Therapeutic antibodies: Smart drugs. Antibodies are highly specific and can therefore bind and affect disease-specific targets, thereby sparing normal cells, and causing fewer toxic side effects than traditional cytotoxic chemotherapies. The immunoglobulin diversity in conjunction with the availability of the technology in monoclonal antibody production, means that antibodies have potential applications as novel diagnostics and therapeutics (Table 13.13), and seem ideal candidates for bioengi-

TABLE 13.13 Some therapeutic monoclonal antibodies

Name (trade name)	Antibody form	Target	Mechanism	Disease	Company
Rituximab (Rituxan)	Chimeric IgG1	CD20	ADCC, CDC, induction of apoptosis	BCL	Genentech, IDEC Pharm.
Trastuzumab (Herceptin)	Humanized IgG1	HER2	Induction of HER2-mediated tumor cell proliferation and migration	Breast	Genentech
Gemtuzumab ozoqamicin (Mytotarg)	Humanized IgG4 linked to calicheamicin	CD33	Delivery of calicheamcin into leukemic cells resulting in DNA strand break and apoptosis	AML	Wyeth-Ayerst, Celltech Group
Alemtuzumab (Campath)	Humanized IgG1	CD52	ADCC, CDC	CLL	Ilex Pharm., Bertex Lab.
Ibritumomab tiuxetan (Zevalin)	Murine IgG1-^{90}Y conjugate	CD20	Delivery of cytotoxic radionuclide, ADCC, CDC, apoptosis	BCL	IDEC Pharm.
Bevacizumab (Avastin)	Humanized IgGi	VEGF	Inhibition of VEGF-induced angiogenesis	Breast, colorectal, renal	Genentech
Epratuzumab (LymphoCide)	Humanized Ig	CD22	Binds and clears CD22 expressing cells	BCL	Immunomedics, Amgen
Pemtumomab (Theragyn)	Murine IgG1-^{90}Y conjugate	PEM	Delivery of toxic radionuclide to PEM-expressing tumor cells	Ovarian	Antisoma, Abbot Labs.
Cetuximab (Erbitux)	Chimeric IgG	EGFR	Binding of EGFR	Colorectal	Bristol-Myers, ImClone
Zamyl (Smart M195)	Humanized IgG1	CD33	ADCC, CDC	AML	Protein Design Labs
Infliximab (Remicade)	Chimeric IgG1	TNFα	Blocking TNFα	RA	Centocor Inc.
Adalimumab (D2E7)	Phage display-derived human Ig	TNFα	Blocking TNFα	RA	Abbot Labs. Cambridge Antib. Technol.
Eculizumab (5G1)	Humanized Ig	C5	Inhibition of C5 cleavage into pro-inflammatory component	RA	Alexion Pharm.
CeaVac	Murine IgG	CA125 antigen	Induction of immune response against tumor-expressed CA125		
OKT-3	Murine IgG	CD3	Organ transplant rejection		Ortho Biotech

Notes: 1. Partially taken from Trikha et *al*. (2002), including those approved by FDA (Food and Drug Administration, U.S.A.) and in development.

2. The suffixes of the generic name of antibodies are: -*omab* for murine antibodies; -x*imab* for chimeric antibodies; -*zumab* for humanized antibodies; and -*umab* for fully human antibodies.

3. Murine antibodies are monoclonal antibodies from mouse. They are extremely immunogenic. Chimeric antibodies are created by replacing mouse constant domains with human constant domains. Additional modifications of framework regions further humanized an antibodiy. Fully human antibodies are derived from human cells or from genetically engineering mice in which murine immunoglobulin genes have been replaced with human antibody genes.

4. Abbreviations used are: ADCC, antibody-dependent cellular cytotoxicity and phagocytosis); AML, acute myelocytic leukemia; BCL, B-cell lymphoma; C5, human complement component; CD, cluster differentiation molecules (surface proteins of different cells); CLL, chronic lymphocytic leukemia; DDC, complement-dependent cytotoxicity; EGFR, epidermal growth factor receptor; HER, human EGF receptor; NHL, non-Hodgkin's lymphoma, PEM, polymorphic epithelial mucine; RA, rheumatoid arthritis; SCLC, small-cell lung cancer; VEGF, vascular endothelial growth factor.

neering to specifically adapt them for these purposes (McCafferty *et al.*, 1996). Therapeutic antibodies can function by three principle modes of action:

1. *Blocking the action of specific molecules*: The blocking activity of therapeutic antibodies is achieved by preventing growth factors, cytokines or other soluble mediators reaching their target receptors, which can be accomplished either by the antibody binding to the factor itself or to its receptor.

2. *Targeting specific cells*: This targeting activity involves directing antibodies toward specific populations of cells and is a versatile approach. Antibodies can be engineered to carry effector moieties, such as enzymes, toxins, radionuclides, cytokines or DNA molecules (gene targeting) to the target cells where the attached moiety can then exert its effect.

3. *Functioning as signaling molecules*: The signaling effect of antibodies is predicated on either inducing cross-linking of receptors that are, in turn, connected to mediators of cell division or programmed cell death, or directing them toward specific receptors to act as agonists for the activation of specific cell populations.

Therapeutic applications include monoclonal antibodies against cancer (Cragg *et al.*, 1999; Trikha *et al.*, 2002), transplant rejection (Berard, 1999) and rheumatoid arthritis (Maini *et al.*, 1999). For example, most oncology antibodies are designed to directly target antigens exposed on tumor cells with cell surface receptors and cluster differentiation (CD) molecules as major targets (e.g. rituxan, trastuzumab, bevacizumab), e.g. CD20 protein is present on the surface of greater than 95% of B-cell lyphomas. Tumor growth is dependent on angiogenesis (formation of new blood vessels), thus targeting tumor vessels is an attractive approach in addition to cytotoxic tumor-directed approaches. Advantages of targeting tumor vasculature are low likelihood of drug resistance, broad application to various tumor types and low toxicity to normal tissues. Targets for monoclonal antibody-based strategies to block angiogenesis include the vascular endothelial growth factor (VEGF), basic fibroblast growth factor (BFGF), epidermal growth factor receptor (EGFR) and cell adhesion molecules (Hicklin *et al.*, 2001).

The most widely explored strategy for enhancing antibody efficacy is the direct conjugation of toxins or radionuclides (e.g. gemtuzumab, ibritumomab). Gemtuzumab ozogamicin is a humanized monoclonal antibody that is linked to the antitumor agent calicheamicin (bacterial toxin). Ibritumomab tiuxetan is a murine antibody attached to ^{90}yttrium (radio-immunotherapeutic antibody) that targets the surface of the mature B cells and B-cell tumors. Both are examples of antibodies designed to specifically deliver their toxic load directly to cancer cells. The use of antibody-directed enzyme prodrug therapy (ADEPT) technology involves the pre-targeting of prodrugs to tumors. An antibody-enzyme fusion protein is first administered and allowed to localize to the tumor site. This is followed by the administration of a prodrug, which is activated by the antibody-enzyme fusion protein at the tumor site to exert the anti-cancer effect (Adams and Weiner, 2005).

Another approach to enhance the effector functions of antibodies is to engineer bispecific antibodies, which comprise two specificities, one for the cell to be eliminated and one for receptors on effector cells (Carter, 2001). Antibody phage display libraries allow for the selection of antibodies of high specificity and affinity toward a variety of different antigens. The antibody fragments, scFv and Fab can do the job normally performed by the intact antibody such as blocking the action of toxins, interactions between cytokines or chemokines with their receptors, binding antigens and carrying effector molecules to their targets. These fragments can be engineered to full antibodies. Coupling the antibody fragments into dimers and trimers generates oligovalent antibody fragments with increased avidity and the ability to cross-link target molecules (Kortt *et al.*, 2001). Diabodies are scFv fragments with two different antigen specificities, one directed against the target and one directed against effector molecules. Other constructs are tandem diabodies and pepbodies (Lunde *et al.*, 2002), which comprise scFv and Fab fragments linked to effector-binding peptide to mimic the complete antibodies binding effector molecules.

13.7.4.4 Catalytic antibody: Abzyme. The similarity of enzymes and immunoglobulins has led to an inquest that the immune system might be exploited to create antibodies with tailored catalytic properties. Abzymes are antibodies engineered to function as specific biocatalysts (section 11.6). These engineering biocatalysts would be especially valuable for those chemical transformations commonly used in the laboratory for which there exist no natural enzyme counterparts. Because many important chemical transformations, including decarboxylations, aldol condensations, S_N2 substitutions, E2 eliminations are sensitive to solvent microenvironment, this strategy of providing suitable microenvironment is likely to be increasingly exploited in the development of a wide variety of catalytic antibodies.

With at least partial understanding of essential features of enzymatic catalysis has come the desire to improve, by the modification of existing enzymes through genetic or chemical means to alter their substrate specificity without loss of catalytic efficiency, the field of protein engineering. A unique feature in the antibody design is that the stereospecific catalysis can be sought for reactions not known to be enzyme catalyzed.

13.8 REFERENCES

Aberham, C., Pendl, C., Gross, P. *et al.* (2001) *Journal of Virology Methods*, **92**, 183–91.

Adams, G. and Weiner, L.M. (2005) *Nature Biotechnology*, **23**, 1147–57.

Aletta, J.M., Cimato, T.R. and Ettinger, M.J. (1998) *Trends in Biochemical Science*, **23**, 89–91.

Alzari, P.M., Lascombe, M.-B. and Poljak, R.J. (1988) *Annual Reviews in Immunology*, **6**, 555–80.

Anderson, J.C., Magliery, T.J. and Schultz, P.G. (2002) *Chemical Biology*, **9**, 237–44.

Ansorge, W., Voss, H. and Zimmermann, J. (1997) *DNA Sequencing Strategies*, John Wiley & Sons, New York.

Arnez, J.G. and Moras, D. (1997) *Trends in Biochemical Sciences*, **22**, 211–6.

Arnheim, N. and Erlich, H. (1992) *Annual Reviews in Biochemistry*, **61**, 131–56.

Arnstein, H.R.V. and Cox, R.A. (1992) *Protein Biosynthesis*, IRL Press/Oxford University Press, Oxford, UK.

Beaudoing, E. and Gautheret, D. (2001) *Genome Research*, **11**, 1520–6.

Bell, S.P. and Dutta, A. (2002) *Annual Reviews in Biochemistry*, **71**, 887–917.

Benkovic, S.J., Valentine, A.M. and Salinas, F. (2001) *Annual Reviews in Biochemistry*, **70**, 181–208.

Berard, J.L. (1999) *Pharmcotherapy*, **19**, 1127–37.

Berger, S.L. and Kimmel, A.R. (eds) (1987) ‘Guide to Molecular Cloning Techniques,’ in *Methods in Enzymology*, Vol. 152, Academic Press, New York.

Blackburn, E.H. (1992) *Annual Reviews in Biochemistry*, **61**, 113–29.

Boue, S., Letunic, I. and Bork, P. (2003) *BioEssays*, **25**, 1031–4.

Bradbury, A.F. and Smyth, D.G. (1991) *Trends in Biochemical Science*, **16**, 112–5.

Brown, T.A. (2001) *Gene Cloning and DNA Analysis*, 4th edn, Blackwell Science, Malden, MA.

Burrell, M.M. (1993) *Enzymes of Molecular Biology*, Humana Press, Totowa, NJ.

Calado, R.T. and Chen, J. (2006) *BioEssays*, **28P**, 109–12.

Campbell, J.A., Davies, G.J., Bulone, V. and Henrissat, B. (1997) *Biochemistry Journal*, **326**, 929–39.

Carter, P. (2001) *Journal of Immunology Methods*, **248**, 7–15.

Casal, J.I. *et al.* (1987) *Biochemistry*, **26**, 1258–64.

Cech, T.R. (2004) *Nature*, **428**, 263–4.

Chaiken, J.M. (1981) *CRC Critrical Reviews in Biochemistry*, **11**, 255–301.

Champoux, J.J. (2001) *Annual Reviews in Biochemistry*, **70**, 369–413.

Cooper, A.A. and Stevens, T.H. (1995) *Trends in Biochemical Sciences*, **20**, 351–6.

Corsi, A.K. and Schekman, R. (1996) *Journal of Biology and Chemistry*, **271**, 30299–302.

Cragg, M.S., French, R.R. and Glennie, M.J. (1999) *Current Opinions in Immunology*, **11**, 541–7.

Craik, C.S., Roczniak, S., Sprang, S. *et al.* (1987) *Journal of Cellular Biochemistry*, **33**, 199–211.

D’Aquino, J.A., Gomez, J., Hilser, V.J. *et al.* (1996) *Proteins*, **25**, 143–56.

Daggett, V. (2002) *Account Chemistry Research*, **35**, 422–9.

Daggett, V. and Fersht, A.R. (2003) *Trends in Biochemical Science*, **28**, 18–25.

Dalbey, R.E. and von Heijne, G. (eds) (2002) *Protein Targeting, Transport and Translocation*, Academic Press, San Diego, CA.

Davies, D.R. and Chacko, S. (1993) *Account Chemistry Research*, **26**, 421–7.

DePamphillis, M.L. (ed.) (1996) *DNA Replication in Eukaryotic Cells*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Deutscher, J. and Saier, M.H. Jr. (1988) *Angrew. Chem. Int. Ed. Engl.*, **27**, 1040–9.

Dobson, C.M., Evans, P.A. and Radford, S.E. (1994) *Trends in Biochemical Science*, **19**, 31–7.

Downward, J. (2004) *British Medical Journal*, **328**, 1245–8.

Draper, D.E. (1996) *Trends in Biochemical Science*, **21**, 145–9.

Drickamer, K. and Taylor, M.E. (1998) *Trends in Biochemical Science*, **23**, 321–4.

Dupureur, C.M., Li, Y. and Tsai, M.-D. (1992) *Journal of the American Chemistry Society*, **114**, 2748–9.

Echols, H. and Goodman, M.F. (1991) *Annual Reviews in Biochemistry*, **60**, 477–511.

Eckstein, F. and Lilley, D.M.J. (eds) (1997) *Mechanisms of Transcription*, Springer, New York.

Ekland, E.H., Szostak, J.W. and Bartel, D.P. (1995) *Science*, **269**, 364–70.

Eun, H.-M. (1996). *Enzymology Primer for Recombinant DNA Technology*, Academic Press, San Diego, CA.

Fedor, M.J. and Uhlenbeck, O.C. (1992) *Biochemistry*, **31**, 12042–54.

Fersht, A.R. (1997) *Current Opinions in Structural Biology*, **7**, 3–9.

Fitzkee, N.C., Fleming, P.J., Gong, H. *et al.* (2005) *Trends in Biochemical Science*, **30**, 73–80.

Fox, T.D. (1987) *Annual Reviews in Genetics*, **21**, 67–91.

Foyer, C.H. (1984) *Photosynthesis*, John Wiley & Sons, New York.

Freedman, R.B., Hirst, T.R. and Tuite, M.F. (1994) *Trends in Biochemical Science*, **19**, 331–6.

Freitag, M. and Selker, E.U. (2005) *Current Opinions in Genetic Development*, **15**, 191–9.

French, D.L., Laskov, R. and Scharff, M.D. (1989) *Science*, **244**, 1152–7.

Friedberg, E.C., Walker, G.C. and Siede, W. (1995) *DNA Repair and Mutagenesis*, ASM Press, Washington, DC.

Frydman, J. (2001) *Trends in Biochemical Science*, **70**, 603–47.

Gahmberg, C.G. and Tolvanen, M. (1996) *Trends in Biochemical Science*, **21**, 308–11.

Gallop, P.M. and Paz, M.A. (1975) *Physics Reviews*, **55**, 418–7.

Gesteland, R.F. and Atkins, J.F. (1996) *Trends in Biochemical Science*, **65**, 741–68.

Gilbert, H.F. (1997) *Journal of Biological Chemistry*, **272**, 29399–402.

Goodman, M.F., Creighton, S., Bloom, L.B. and Petruska, J. (1993) *Critical Reviews in Biochemical Molecular Biology*, **28**, 603–33.

Green, R. and Noller, H.F. (1997) *Annual Reviews in Biochemistry*, **66**, 679–716.

Greene, J.J. and Rao, V.B. (eds) (1998) *Recombinant DNA Principles and Methodologies*, Marcel Dekker, New York.

Grosjean, H. and Benne, R. (eds) (1998) *Modification and Editing of RNA*. ASM Press, Washington, DC.

Haebel, P.W., Gutmann, S. and Ban, N. (2004) *Current Opinions in Structural Biology*, **14**, 58–65.

Harding, J.J. and Crabbe, M.C.J. (eds) (1992) *Post-Translational Modifications of Proteins*. CRC Press, Boca Raton, FL.

Hartley, C.G. and Villeponteau, B. (1995) *Current Opinions in Genetic Development*, **5**, 249–55.

Hayaishi, O. and Ueda, K. (1977) *Annual Reviews in Biochemistry*, **46**, 95–116.

He, L. and Hannon, G.J. (2004) *National Reviews in Genetics*, **5**, 522–31.

Heid, C. (1996) *Genome Research*, **6**, 986–94.

Hendrick, J.P. and Hartl, F.U. (1993) *Annual Reviews in Biochemistry*, **62**, 349–84.

Herschlag, D. and Cech, T.R. (1990) *Biochemistry*, **29**, 10159–71.

Hicklin, D.J., Witte, L., Zhu, Z. *et al.* (2001) *Drug Discovery Today*, **6**, 517–28.

Hohsaka, T. and Sisido, M. (2002) *Current Opinions in Chemical Biology*, **6**, 809–15.

Innis, M., Gelfand, D. and Sninsky, J. (eds) (1999) *PCR Methods Manual*, Academic Press, San Diego.

Jackson, D.Y., Burnier, J., Quan, C. *et al.* (1994) *Science*, **266**, 243–7.

Johnson, A. and O'Donnell, M. (2005) *Annual Reviews in Biochemistry*, **74**, 283–315.

Joyce, C.M. and Steitz, T.A. (1987) *Trends in Biochemical Science*, **12**, 288–92.

Joyce, C.M. and Steitz, T.A. (1994) *Annual Reviews in Biochemstry*, **63**, 777–822.

Kay, B., Winter, J. and McCafferty, J. (1996) *Phage Display of Peptides and Proteins: A Laboratory Manual*, Academic Press, London.

Kim, P.S. and Baldwin, R.L. (1990) *Annual Reviews in Biochemstry*, **59**, 631–60.

Kingdon, H.S., Shapiro, B.M. and Stadtman, E.R. (1967) *Proceedings of the National Academy of Sciences, USA*, **58**, 1703–10.

Kornberg, A. (1988) *Journal of Biology and Chemistry*, **263**, 1–4.

Kornberg, A. and Baker, T.A. (1992) *DNA Replication*, 2nd edn, W.H. Freeman, San Francisco, CA.

Kornfeld, R. and Kornfeld, S. (1985) *Annual Reviews in Biochemstry*, **54**, 631–64.

Kortt, A.A., Dolezal, O., Power, E.B. and Hudson, P.J. (2001) *Biomolecular Engineering*, **18**, 95–108.

Ledl, F. and Schleicher, E. (1990) *Angew. Chem. Intl. Ed., Engl.*, **29**, 565–706.

Lee, D. and Lorsch, J.R. (2004) *Annual Reviews in Biochemstry*, **73**, 657–704.

Lev-Maor, G., Sorek, R., Shomron, N. and Ast, G. (2003) *Science*, **300**, 1288–91.

Lewin, B. (1987) *Genes*, 3rd edn, John Wiley & Sons, New York.

Lindahl, T. and Barnes, D.E. (1992) *Annual Reviews in Biochemstry*, **61**, 251–81.

Lindorff-Larsen, K., Vendruscolo, M., Paci, E. and Dobson, C.M. (2004) *Natural Structural Molecular Biology*, **11**, 443–9.

Lindorff-Larsne, K., Rogen, P., Paci, E. *et al.* (2005) *Trends in Biochemical Science*, **30**, 13–9.

Lohman, T.M. and Bjornson, K.P. (1996) *Annual Reviews in Biochemstry*, **65**, 169–214.

Lu, Q. and Weiner, M.P. (eds) (2001) *Cloning and Expression Vectors for Gene Function Analysis*, Eaton Publishers, Natick, MA.

Lunde, E., Lauvrak, V., Rasmussen, I.B., Schjetne, K.W. *et al.* (2002) *Biochemistry Society Transactions*, **30**, 500–6.

Ly, H.D. and Withers, S.G. (1999) *Annual Reviews in Biochemstry*, **68**, 487–522.

Mrabet, N.T. *et al.* (1992) *Biochemistry*, **31**, 2239–53.

Maini, R., Clair, E.W., Breedveld, F. *et al.* (1999) *Lancet*, **35**, 1932–9.

Mandal, M. and Breaker, R.R. (2004) *National Reviews in Molecular Cellular Biology*, **5**, 451–63.

Maquat, L.E. (2004) *National Reviews in Molecular Cellular Biology*, **5**, 89–99.

Marians, K.J. (1992) *Annual Reviews in Biochemstry*, **61**, 673–719.

Matson, S.W. and Kaiser-Rogers, K.A. (1990) *Annual Reviews in Biochemstry*, **59**, 289–329.

Matthews, B.W. (1995) *Advances in Protein Chemistry*, **46**, 249–78.

McCafferty, J., Hoogenboom, H.R. and Chiswell, D.J. (1996) *Antibody Engineering: A Practical Approach*, IRL Press. Oxford, UK.

McIlhinney, R.A.J. (1990) *Trends in Biochemical Science*, **15**, 387–91.

McKnight, S.L. and Yamamoto, K.R. (eds) (1992) *Transcriptional Regulation*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

McManus, M.T. and Sharp, P.A. (2002) *Nature Reviews in Genetics*, **3**, 737–47.

McPherson, M.J. (1991) *Directed Mutagenesis*, IRL Press, Oxford, UK.

Meister, G. and Tuschl, T. (2004) *Nature*, **431**, 343–9.

Mikheeva, S. and Jarrell, K.A. (1996) *Proceedings of the National Academy of Sciences, USA*, **93**, 7486–90.

Moldave, K. (ed.) (1981) *RNA and Protein Synthesis*, Academic Press, New York.

Muir, T.W. (2003) *Annual Reviews in Biochemstry*, **72**, 249–89.

Mullis, K.B., Ferré, F. and Gibbs, R.A. (1994) *The Polymerase Chain Reaction*, Birkhaüser, Boston, MA.

Muraki, M., Harata, K. and Jigami, Y. (1992) *Biochemistry*, **31**, 9212–9.

Natsuka, S. and Lowe, J.B. (1994) *Current Opinions in Structural Biology*, **4**, 683–91.

Newton, C.R. and Graham, A. (1997) *PCR*, BIOS Scientific Publishers, Oxford, UK.

Ng, H.H. and Bird, A. (2000) *Trends in Biochemical Science*, **25**, 121–6.

Niemeyer, C.M., Adler, M. and Wacker, R. (2005) *Trends in Biotechnology*, **23**, 208–16.

Ogawa, T. and Okazaki, T. (1980) *Annual Reviews in Biochemstry*, **49**, 421–57.

Ogle, J.M., Carter, A.P. and Ramakrishnan, V. (2003) *Trends in Biochemical Science*, **28**, 259–66.

Omoto, C.K. and Lurquin, P.F. (2004) *Genes and DNA: A Beginners Guide to Genetics and its Applications*, Columbia University Press, New York.

Onoa, B. and Tinoco, I. Jr. (2004) *Current Opinions in Structural Biology*, **14**, 374–9.

Orlando, C., Pinzani, P. and Pazzagli, M. (1998) *Clinical Chemistry Laboratory Methods*, **36**, 255–69.

Padgett, R.A., Grabowski, P.J. and Knonarska, M.M. (1986) *Annual Reviews in Biochemstry*, **55**, 1123–50.

Paik, W.K. and Kim, S. (1975) *Advances in Enzymology*, **42**, 227–86.

Paulus, H. (2000) *Annual Reviews in Biochemstry*, **69**, 447–496.

Perler, F.B., Olsen, G.J. and Adam, E. (1997) *Nucleic Acids Research*, **25**, 1087–93.

Pingoud, A.M. (ed.) (2004) *Restriction Endonucleases*, Springer-Verlag, New York.

Pyle, A.M. and Green, J.B. (1994) *Biochemistry*, **33**, 2716–25.

Quéméneur, E., Moutiez, M., Charbonnier, J.P. and Ménez, A. (1998) *Nature*, **391**, 301.

Radr, C. and Barbas, C.F. (1997) *Current Opinions in Biotechnology*, **8**, 503–8.

Rajeevan, M.S., Vernon, S.D., Tayavang, N. and Unger, E.R. (2001) *Journal of Molecular Diagnosis*, **3**, 26–31.

Ramakrishnana, V. (2002) *Cell*, **108**, 557–72.

Rapoport, T.A., Jungnickel, B. and Kutay, U. (1996) *Annual Reviews in Biochemstry*, **65**, 271–303.

Rio, D.C. (1992) *Current Opinions in Cell Biology*, **4**, 444–52.

Roberts, R.J., Vincze, T., Posfai, J. and Macelis, D. (2005) *Nucleic Acids Research*, **33**, D230–2.

Rodnian, M.V. and Wintermeyer, W. (2001) *Annual Reviews in Biochemstry*, **70**, 415–35.

Roeder, R.G. (1996) *Trends in Biochemical Science*, **21**, 327–35.

Salas, M. (1991) *Annual Reviews in Biochemstry*, **60**, 39–71.

Sambrook, J., Fritsch, E.F. and Maniatas, T. (eds) (1989) *Molecular Cloning, A Laboratory Manual*, 2nd edn, Cold Spring Harbor Laboratory Press, New York.

Seeberg, E., Eide, L. and Bjørås, M. (1995) *Trends in Biochemical Science*, **20**, 391–7.

Shakhnovich, E.I., Broglia, R.A. and Tiana, G. (eds) (2001) *Protein Folding, Evolution and Design*, IOS Press, Washington, DC.

Sigler, P.B., Xu, Z., Rye, H.S. *et al.* (1998) *Annual Reviews in Biochemstry*, **67**, 581–608.

Skalka, A.M. and Goff, S.P. (eds) (1993) *Reverse Transcriptase*, Cold Spring Harbor Press, Cold Spring Harbor Lab. New York.

Smith, M. (1985) *Annual Reviews in Genetics*, **19**, 423–62.

Smotrys, J.E. and Linder, M.E. (2004) *Annual Reviews in Biochemstry*, **73**, 559–87.

Stamm, S., Riethoven, J.J., LeTexier, V., Gapalakrishnan, C., Kumaduri, V., Tang, V., Barbosa-Morais, N.L. and Tangavel, A. (2006) *Nucleic Acid Resear*, **34**, D46–D55.

Stebbins, J.W. and Kantrowitz, E.R. (1992) *Biochemistry*, **31**, 2328–32.

Steiner, D.F., Smeekens, S.P., Ohagi, S. and Chan, S.J. (1992) *Journal of Biology and Chemistry*, **267**, 23435–8.

Steitz, T.A. (1993) *Current Opinions in Structural Biology*, **3**, 31–8.

Steitz, T.A. (1998) *Nature*, **391**, 231–2.

Stemmer, W.P.C. (1994) *Nature*, **370**, 389–91.

Stoddart, R.W. (1984) *The Biosynthesis of Polysaccharides*, MacMillan Publishing Co., New York.

Suttie, J.W. (1980) *Trends in Biochemical Science*, **5**, 302–5.

Tang, G. (2005) *Trends in Biochemical Science*, **30**, 106–14.

Tartakoff, A.M. and Singh, N. (1992) *Trends in Biochemical Science*, **17**, 470–3.

Tonegawa, S. (1988) *Angew. Chem. Intl. Ed. Engl.*, **27**, 1028–39.

Trikha, M., Yan, L. and Nakada, M.T. (2002) *Current Opinions in Biotechnology*, **13**, 609–14.

Wade, P.A., Pruss, D. and Wolffe, A.P. (1997) *Trends in Biochemical Science*, **22**, 128–32.

Waga, S. and Stillman, B. (1998) *Annual Reviews in Biochemstry*, **67**, 721–51.

Walker, G.C. (1995) *Trends in Biochemical Science*, **20**, 416–20.

Wang, J.C. (1996) *Annual Reviews in Biochemstry*, **65**, 635–92.

Watson, J.D., Gilman, M., Witkowski, J. and Zoller, M. (1992) *Recombinant DNA*, 2nd edn, Scientific American/W.H. Freeman, New York.

Weiss, R.B., Huang, W.M. and Dunn, D.M. (1990) *Cell*, **62**, 117–26.

Wilks, H.M. *et al.* (1990) *Biochemistry*, **29**, 8587–91.

Wilks, H.M. *et al.* (1988) *Science*, **242**, 1541–4.

Winkler, W.C., Nahvi, A. and Breaker, R.R. (2002) *Nature*, **419**, 952–6.

Winkler, W.C., Nahvi, A., Roth, A. *et al.* (2004) *Nature*, **428**, 281–6.

Withey, J.H. and Friedman, D.I. (2003) *Annual Reviews in Microbiology*, **57**, 101–23.

Wold, F. (1981) *Annual Reviews in Biochemstry*, **50**, 783–841.

Yaoi *et al.* (1996) *Journal of Biochemistry*, **119**, 1014–8.

Yelin, R., Dahary, D. and Sorek, R. (2003) *Nature Biotechnology*, **21**, 379–86.

Young, B., Herschlag, D. and Cech, T.R. (1991) *Cell*, **67**, 1007–19.

Yuan, R. (1981) *Annual Reviews in Biochemstry*, **50**, 285–315.

Zhang, A.J. and Cech, T.R (1986) *Science*, **231**, 470–5.

Zhang, F.L. and Casey, P.J. (1996) *Annual Reviews in Biochemstry*, **65**, 241–69.

Zhang, A.J., Davila-Aponte, J.A. and Cech, T.R. (1994) *Biochemistry*, **33**, 14935–47.

World Wide Webs cited

American Type Culture Collection:	http://www.atcc.org/
Antisense RNADB:	http://bio-inf.net/aobase
ASD:	http://www.ebi.ac.uk/asd
Codon usage from GenBank (CUTG):	http://www.kazusa.or.jp/codon/
Eukaryotic promoter database (EPD)	http://www.epd-isb-sib.ch
Functional molecular immunology:	http://research.i2r.a-star.edu/sg/fimm
Genetic code:	http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi
Human major histocompatibity complexes DB	http://www.ncbi.nlm.nih.gov/mhc/
Haptendb:	http://imtech.res.in/raghava/haptendb/
IMGT:	http://imgt.cines.fr/
microRNA Registry:	http://www.sanger.ac.uk/Software/Rfam/mirna/
MPDB:	http://www.biotech.ist.unige.it/interlab/mpdb.html
PlasMapper:	http://wishart.biology.ualberta.ca/PlasMapper/
PolyA_DB:	http://polya.umdnj.edu/polyadb
Pimer3:	http://genome.wi.mit.edu/cgi-bin/primer/primer3wwwcgi
PrediSi	http://www.predisi.de
PROBEmer:	http://probemer.cs.loyola.edu/
REBASE:	http://rebase.neb.com/rebase/rebase.html
siRNAdb:	http://sirna.cgb.ki.se/
TRANFAC:	http://transfac.gbf.de/TRANFAC/cl/cl.html
VectorDB:	http://genome-www2.stanford.edu/vectordb/
Web Primer:	http://genome-www2.stanford.edu/egi-n/SGD/web-primer

CHAPTER 14

BIOMACROMOLECULAR INFORMATICS

14.1 OVERVIEW

Bioinformatics is the science of using information technology to understand biology. It refers to the use of computers to characterize the molecular components of living organisms and other biological data of which biomacromolecular informatics is the central concern (Attwood and Parry-Smith, 1999; Baxevanis and Ouellette, 2005; Lesk, 2005; Ramsden, 2004), because the major components of bioinformatics consist of genomics, proteomics and glycomics. The science of informatics is concerned with the representation, organization, manipulation, distribution, maintenance and use of information. Thus bioinformatics is a tool with the objective of elucidating how living systems work dealing with collection, representation, management, analysis and distribution of biological data (Figure 14.1).

Genomics is the study of the complete genomes including nuclear and extranuclear genes of an organism (Singer and Berg, 1991). Proteomics investigates the complete protein complements of genomes (Liebler, 2002). Glycomics then studies the complete glycan complement of the organism. In metabolomics, the full complement of metabolites of cell, tissue or organism is studied. Structural studies of cellular DNA (genomics), proteins (proteomics) and glycans (glycomics) focus on analyses and interpretations of sequences as well as 3D structures of biomacromolecules. Functional studies have emphasized analyses of gene expression (transcriptomics), protein translation (proteomics) including posttranslational modifications, and the metabolic network (metabolomics) with a view to a system biology approach toward molecular definition of the phenotypes of biological systems. Thus biosequences, 3D structural and functional data/information are collected. Methodologies for acquisition and collection of these data have been described in previous chapters and only a few will be described here. This chapter will introduce computer technology that manages, analyzes, archives and retrieves biodata/information.

14.2 BIOSEQUENCES

14.2.1 Sequencing biomacromolecules

Often preparatory isolation, purification and identification of the biomacromolecules of interest are performed prior to the sequence data acquisition/collection. Some of tools to isolate and purify these biomacromolecules have been described in Chapter 3. The basic strategy of sequencing biomacromolecules (Brown, 1994; Durbin, *et al*., 1998; Findley and Geisow, 1998) is summarized in Table 14.1, as follows:

Biomacromolecules, by C. Stan Tsai
Copyright © 2007 John Wiley & Sons, Inc.

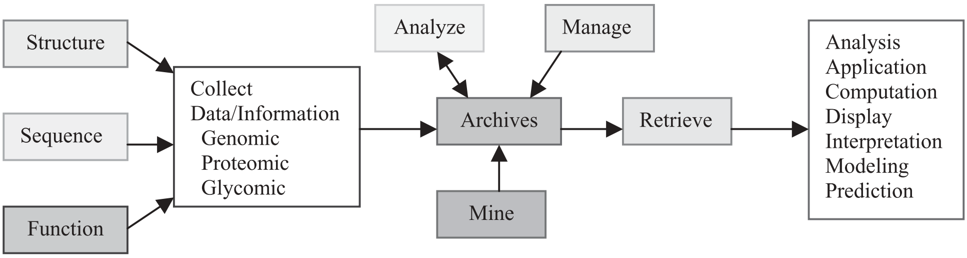


Figure 14.1 Components of biomacromolecular informatics and their interactions

TABLE 14.1 Sequencing strategies

	Genomics	Proteomics	Glycomics
Target	DNA, RNA	Proteins	Glycans of glyco-proteins/lipids
Major sequence archives	NCBI, EBI/EMBL, DDBJ	EBI, PIR-PSD, Swiss-Prot, UniProt	Glycan Database, SweetDB
Cleavage/fragmentation	Restriction endonucleases	Specific proteases	Acid hydrolysis
Fragment separation	hrPAGE	PAGE, HPLC, 2DE	HPLC, lectin affinity
Terminal sequencing	Dideoxy chain termination 5′→3′	Edman degradation N→C	Exoglycosidases Nonreducing→reducing
Internal sequencing	Hybridization, MS	MS, NMR, ladder sequcencing	Chemical, MS, NMR, endoglycosidases.
Informatics	Submission/deposition of sequence data to the above sequence archives		
Reference, subsection	4.2	5.2	6.2

Notes: Abbreviations (Web sites) used are, EBI (http://www.ebi.ac.uk), European Bioinformatics Institute of EMBL (European Molecular Biology Laboratories at http://www.embl-heidelberg.de) data library, DDBJ (http://www.ddbj.nig.ac.jp), DNA Data Bank of Japan; NCBI (http://www.ncbi.nlm.nih.gov), National Center for Biotechnology Information; PIR (http://pir.georgetown.edu), Protein Information Resources; Swiss-Prot (http://www.expasy.org/sprot/), the Swiss Institute of Bioinformatics in collaboration with EMBL; UniProt (http://www.uniprot.org/), Glycan Database (http://www.functionalglycomics.org/), SweetDB (http://www.glycosciences.de/sweetdb/index.php.

1. *Search database*: Using information available at hand, such as keyword (e.g. name, function if known), biological source, molecular weight, composition, etc to search databases for the identity (sequence is already known) and homology (related sequences exist).
2. *The specific and reproducible cleavage of the biomacromolecular chain into fragments of manageable size*: Enzymes that cleave biopolymeric bonds at specific residues are used to produce fragments.
2. *Fractionation of the resulting fragments*: Fragments are commonly fractionated by chromatographic and/or electrophoretic separations.
3. *Sequencing of individual fragments*: The sequencing can begin at one of the termini in a terminal sequencing experiment or can be conducted within the biopolymer chain in an internal sequencing experiment.
4. Repetition of the preceding steps using a cleavage procedure that yields a set of the fragments, which overlap the cleavage sites in the previous set(s).
5. The ordering of the completely sequenced fragments using the information from the overlapped cleavage sites.

14.2.2 Sequence similarity and pair-wise alignment

When the sequences of genes and proteins are compared, the similarities and differences at the level of individual bases or amino acids are analyzed with the aim of inferring structural, functional and evolutionary relationships among the sequences under study. The most common comparative method is sequence alignment, which provides an explicit mapping between the residues of two or more sequences. A sequence consists of letters selected from an alphabet. The complexity of the alphabet is defined as the number of different letters it contains. The complexity is 4 for nucleic acids, 20 for proteins and varied (commonly 18, Table 6.1) for polysaccharides. Sometimes additional characters are used to indicate ambiguities in the identity of a particular base or residue, for example B for Asp or Asn and Z for Glu or Gln in proteins and R for A or G, and Y for C, T or U in nucleic acids. The computational juxtaposition of two or more linear strings of nucleotides or amino acids is thus known as a sequence alignment. The practice of aligning sequences of nucleic acids and proteins is based on the assumption that linear similarity maps with functional similarity because gene information is converted into protein information in a linear manner according to the genetic code. Proteins that share similar amino acid sequences fold similarly and therefore function in a similar manner.

14.2.2.1 The evolutionary basis of sequence alignment. The term, similarity is an observable quantity that might be expressed as a suitable measure, such as percent identity. Homology, on the other hand, refers to a conclusion drawn from these data that two genes share a common evolutionary history. Thus sequences are homologous if they are related by divergence from a common ancestor. Among homologous sequences, it is useful to distinguish between biomacromolecules that perform the same function in different species (these are referred to as orthologues) and those that perform different but related functions within one organism (so-*called paralogues*). Sequence comparison of orthologous nucleic acids/proteins opens the way to the study of molecular paleontology. In particular cases, construction of phylogenetic trees has revealed relationships among different species. The study of paralogous nucleic acids/proteins, however, has provided deeper insight into the underlying mechanisms of evolution. Paralogous proteins arose from single genes via successive duplication events. The duplicated genes have followed separate evolutionary pathways, and new specificity has evolved through variation and adaptation. In addition, many proteins may converge to folding arrangements that have particularly favorable packing and proteins related in this manner are generally referred to as analogues (Russell *et al*., 1997).

While it is presumed that homologous sequences have diverged from a common ancestral sequence through iterative molecular changes, we do not actually know what the ancestral sequence was. All we have to observe are the sequences from extant organisms. The changes that occur during divergence from the common ancestor can be characterized as substitutions, insertions and deletions. Residues that have been aligned, yet are not identical, would represent substitutions. Regions in which the residues of one sequence correspond to nothing in the other would be interpreted as either an insertion into one sequence or a deletion from the other (InDel). These gaps are usually represented in the alignment with consecutive dashes aligned with the residues. Apart from strict identity in pattern matching, one approach in the sequence comparison of proteins can be introduced by considering amino acid residues as members of groups with shared biochemical properties (Table 14.2).

In a residue-by-residue alignment it is often apparent that certain regions of a nucleic acid/protein (or specific nucleotides/amino acids) are more highly conserved than others.

TABLE 14.2 Classification of amino acids according to their biochemical properties

Property	Residue
Small	Ala, Gly
Hydroxyl/phenol	Ser, Thr, Tyr
Thiol/thioether	Cys, Met
Acidic/amide	Asp, Asn, Gln, Glu
Basic	Arg, His, Lys
Polar	Ala, Cys, Gly, Pro, Ser, Thr
Aromatic	Phe, Trp, Tyr
Alkyl hydrophobic	Ile, Leu, Met, Val

This information may be suggestive of which residues are most crucial in maintaining a nucleic acid/protein's structure or function. However, there may be other positions that do not play a significant functional role yet happen to be identical for historical reasons if the sequences are taken from closely related species. Nevertheless, sequence alignment provides a useful way to gain new insights using existing knowledge to deduce structural and functional properties of a novel protein from comparison to those that have been well studied. Upon observing a surprisingly high degree of sequence similarity between two genes or proteins, we might infer that they share a common evolutionary history and from this we might anticipate that they have similar biological functions. It must be emphasized, however, that these inferences should not be assumed to be correct based on computational analysis alone; they must always be tested experimentally.

Using sequence analysis techniques, it is feasible to identify similarities between novel query sequences and database sequences whose structures and functions have been elucidated. This is straightforward at high levels of sequence identity (above 50% identity), where relationships are clear, but at low levels (below 50% identity) it becomes increasingly difficult to establish relationships reliably. Significant false positives occur in the twilight zone for 20–30% sequence identity with unreliable results in the midnight zone (below 20% identity). Alignment of random sequences can produce around 20% identity; below that the alignments are no longer statistically significant. In the twilight zone, relationship recognition can be improved by building sequence profiles using a set of known homologues from a protein family to determine the constraints. In the midnight zone, it is often necessary to exploit structure data in determining the relationships. Pairwise comparison of two sequences is a fundamental process in sequence analysis. It defines the concepts of sequence identity, similarity and homology as applied to two proteins, DNA or RNA sequences. Homology is not a measure of similarity, but a statement that sequences have a divergent rather than a convergent relationship. Database interrogation can take the form of text queries or sequence similarity searches. Typically, the user employs a query sequence to conduct sequence similarity search so that the relationships between the query sequence (probe) and another sequence (target) can be quantified and their similarity assessed.

14.2.2.2 Dot matrix representation in sequence comparison. The dot matrix representation (dotplot) is a simple picture that gives an overview of the similarities between two sequences. It is a matrix with the row corresponding to the residues of one sequence and the column to the residues of the other sequence.

In dot matrix representations, the basic idea is to use the sequences as the coordinates of a two-dimensional graph and then plot points of correspondence within the inte-

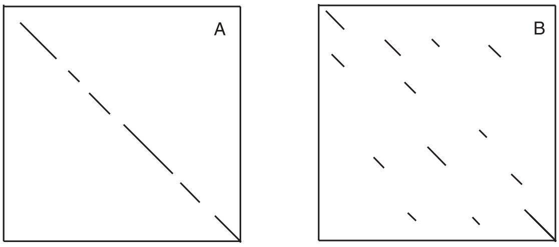


Figure 14.2 Dotplot for sequence comparison

rior. In its simplest form, the positions in the dotplot are left blank if the residues are different, and filled if they match. Each dot usually indicates that within some small window, the sequence similarity is above some cutoff. When two sequences are consistently matching over an extended region, the dots will merge to form a diagonal line segment (Figure 14.2A). Often regions of similarity may be displaced, to appear on parallel but not colinear diagonals (Figure 14.2B). This indicates that insertions or deletions have occurred in the segments between the similar regions. The dotplot can be generated using the DOTTER program (ftp://ftp.sanger.ac.uk/pub/dotter/). The program relies on finding the optimal path that represents the alignment. The Web resource for dotplots can be accessed at http://www.cgr.ki.se/cgr/groups/sonnhammer/Dotter.html

14.2.2.3 Similarity scoring. A comprehensive alignment must account fully for the positions of all residues in the sequences under comparison. To achieve this, gaps may be inserted so as to maximize the number of identical matches. However, the result of such a process is biologically meaningless. Thus scoring penalties are introduced to minimize the number of gaps that are opened and extension penalties are also incurred when a gap has to be extended. The total alignment score is a function of the identity between aligned residues and the incurred gap penalties. By doing this, only residue identities are considered, resulting in a unitary matrix that weights identical residue matches with a score of 1 and other elements being zero. In order to improve diagnostic performance, the scoring potential of the biologically significant signals is enhanced so that they can contribute to the matching process without amplifying noise. A need to balance between high-scoring matches that have only mathematical significance and lower-scorings that are biologically meaningful is the essential requisite of sequence analysis. Scoring matrices that weight matches between nonidentical residues based on evolutionary substitution rates have been devised to address this need. Such tools may increase the sensitivity of the alignment, especially in a situation where sequence identity is low. One of the most popular scoring matrices is Dayhoff mutation data (MD) matrix (Dayhoff *et al.*, 1978).

The MD score is based on the concept of the point accepted mutation (PAM). Thus one PAM is a unit of evolutionary divergence in which 1% of the amino acids have been changed. If these changes were purely random, the frequencies of each possible substitution would be determined simply by the overall frequencies of the different amino acid (background frequencies). However, in related proteins, the observed substitution frequencies (target frequencies) are biased toward those that do not seriously disrupt the protein's function. In other words these are mutations that have been accepted during evolution. Thus the substitution scores in the mutation probability matrix are then proportional to the natural log of the ratio of target frequencies to background frequencies. Mutation probability matrices corresponding to longer evolutionary distance can be derived by multiplication of this matrix of probability values by itself the appropriate number of times (n times for a distance of n PAM). Table 14.3 shows the values

TABLE 14.3 PAM Scale for amino acid residues

Observed % difference	Evolutionary distance in PAM
1	1
3	5
10	11
15	17
20	23
25	30
30	38
40	56
50	80
60	112
70	159
80	246

connecting the overall proportion of observed mismatches in amino acid residues with the corresponding evolutionary distance in PAM (PAM scale).

The 250 PAM (250 accepted point mutations per 100 residues) matrix gives similarity scores equivalent to 80% difference between two protein sequences. Because the sequence analysis is aimed at identifying relationships in the twilight zone, the MD for 250 PAM has become the default matrix in many analysis packages. The appropriate similarity score for a particular pair-wise comparison at any given evolutionary distance is the logarithm of the odds that a particular pair of amino acid residues has arisen by mutation. These odds can be derived via dividing the values in the mutation probability matrix by the overall frequencies of the amino acid residues. In the matrix, residues likely to undergo mutation, neutral (random mutations), and those unlikely to undergo mutation have values of greater than 0, equal to 0, and less than 0 respectively.

Most sequence alignment methods seek to optimize the criterion of similarity. There are two approaches of sequence alignments; a global alignment compares similarity across the full stretch of sequences while a local alignment searches for regions of similarity in parts of the sequences.

14.2.2.4 Global alignment. Global pair-wise alignment makes comparison over the entire lengths of the two sequences. The approach (Needleman and Wunsch, 1970) can be considered as proceeding through four basic steps: laying out the alignment matrix, initializing the matrix, 'wave front' updating the matrix elements, and the trace back. Accordingly, a 2D matrix is constructed by comparing two sequences that are placed along the *x*- and the *y*-axes respectively (Table 14.4). For initiation, cells representing identities are scored 1, and those with mismatches are scored 0 to populate the 2D array with 0's and 1's.

For the wave-front update of the matrix, each cell in the matrix is assigned a new value. An operation of successive summation of cells begins at the last cell (e.g. S,T) in the matrix progressing to the next raw or column by adding the maximum number from two constituent sub-paths to the cell. For the cell (e.g. R,R) in the example, the original score is 1. The maximum value of the two sub-paths is 5, thus the new value for the cell (e.g. R,R) becomes 5 + 1 = 6. This process is repeated for every cell in the matrix (Table 14.5).

An alignment is generated, by working through (trace back) the completed matrix, starting at the highest-scoring element at the N-terminal and following the path of high scores through to the C-terminal. For example:

TABLE 14.4 Unitary scoring matrix

	C	H	E	M	I	C	A	L	R	E	A	G	E	N	T
C	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0
H	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0
M	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
R	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
G	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
E	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

TABLE 14.5 Completed maximum-match matrix

	C	H	E	M	I	C	A	L	R	E	A	G	E	N	T
C	11	9	8	7	6	7	6	6	6	5	4	3	2	1	0
H	9	10	8	7	6	6	6	6	6	5	4	3	2	1	0
E	8	8	9	7	6	6	6	6	5	6	4	3	2	1	0
M	7	7	7	8	6	6	6	6	5	5	4	3	2	1	0
I	6	6	6	6	7	6	6	6	5	5	4	3	2	1	0
S	6	6	6	6	6	6	6	6	5	5	4	3	2	1	0
T	6	6	6	6	6	5	6	6	5	5	4	3	2	1	1
R	5	5	5	5	5	5	5	5	*6*	5	4	3	2	1	0
Y	5	5	5	5	5	5	5	5	5	*5*	*4*	*3*	*2*	*1*	*0*
A	4	4	4	4	4	4	5	4	4	*4*	5	3	2	1	0
G	3	3	3	3	3	3	3	3	3	*3*	3	4	2	1	0
E	2	2	3	2	2	2	2	2	2	*3*	2	2	3	1	0
N	1	1	1	1	1	1	1	1	1	*1*	1	1	1	2	0
T	0	0	0	9	0	0	0	0	0	*0*	0	0	0	0	1
S	0	0	0	0	0	0	0	0	0	*0*	0	0	0	0	0

```
CHEMICALREAGENT-
||| ||||||||||||
CHEMIST-RYAGENTS
```

14.2.2.5 Local alignment. Two sequences that are distantly related to each other may exhibit small regions of local similarity, though no satisfactory overall alignment can be found. The Smith and Waterman algorithm (Smith and Waterman, 1981) is designed to find these common regions of similarity and has been used as the basis for many subsequent algorithms.

The FASTA (Lipman and Pearson, 1985) and Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, 1990; 1997) programs are local similarity search methods that concentrate on finding short identical matches, which may contribute to a total match.

FASTA uses *ktups* (1 or 2 amino acids for proteins and 4 or 6 nucleotides for nucleic acids) while BLAST uses words (3 amino acids for proteins and 11 nucleotides for nucleic acids) of short sequences in the database searching. In a typical output from a FASTA search, the name and version of the program are given at the top of the file with the appropriate citation. The query sequence together with the name and version of the search database are indicated. A range of parameters used by the algorithm and the run-time of the program are printed. Following these statistics come the results of the database search with a list of a user-defined number of matches with the query sequence and information for the source database (identifier, ID code, accession number and title of the matched sequences). The length in amino acid residues of each of the retrieved matches is indicated in brackets. Various initial and optimized scores and E(expected)-values are given. After the search summary, the output presents the complete pair-wise alignments of a user-defined number of hits with the query sequence. Within the alignments, identities are indicated by the ':' character, while similarity is indicated by the '.' character.

BLAST (Altschul *et al*., 1990) searches two sequences for sub-sequences of the same length that form an ungapped alignment. The algorithm creates a list of all short sequences (BLAST words) that score above a threshold value when aligned with the query sequence. The sequence database is searched for occurrences of these words and the matching words are then extended into ungapped local alignments between the query sequence and the sequence from the database. The resulting high scoring pairs (HSPs) form the basis of the ungapped alignments that characterize BLAST output. In the gapped BLAST (Altschul *et al*., 1997), the algorithm seeks only one ungapped alignment that makes up a significant match. Dynamic programming is used to extend a central pair of aligned residues in both directions to yield the final gapped alignment to speed up the initial database search. In the typical BLAST output, the name and version of the program together with the appropriate citation are given at the top of the file. The query sequence and the name of the search database are then indicated. Next lines start with the results of the database search consisting of a list of a user-defined number of matches with the query sequence and information for the source database (identifier, ID code, accession number and title of the matched sequences). The highest score of the set of matching segment pairs for the given sequence with the number of HSPs denoted by the parameter N is given. This is followed by a p(probability)-value. After the search summary, the output presents the pair-wise alignments of the HSPs for each of the user-defined number of hits with the query sequence. For each aligned HSP, the beginning and end locations within the sequence are marked and identities between them are indicated by the corresponding nucleotide/amino acid symbols. BLAST has its greatest utility in sequence database searching and retrieval.

14.2.3 Similarity search and multiple sequence alignment

Searching in databases for homologues of known DNA/proteins is a central theme of bioinformatics. The goals of database searching are high sensitivity (i.e. picking up even very distant relationships) and high selectivity (i.e. minimizing the number of sequences reported that are not true homologues). This is accomplished by multiple sequence alignments. The objective of multiple sequence alignment is to generate a concise summary of sequence data in order to make an informed decision on the relatedness of sequences to a gene family. Alignments should be regarded as models that can be used to test hypothesis. There are two main approaches on the construction of alignments. The first method is guided by the comparison of similar strings of nucleotide/amino acid residues taking into account their physicochemical properties, mutability data, etc. The second method results from comparison at the level of secondary or tertiary structure where alignment positions are determined solely on the basis of structural equivalence. Alignments generated from these different approaches often show significant disparities. Each is a model that reflects a different biochemical view of sequence data. The multiple sequence analysis draws together related sequences from various species and expresses the degree of similarity in a relatively concise format.

There are two basic search strategies for finding a set of related sequences, namely keywords and similarity. A keyword search identifies sequences by looking through their written descriptions (i.e. the annotation section of a database file) and a similarity search looks at the sequences themselves. The keyword search is easier and seems more intuitive, but it is far from exhaustive. The main search engines for similarity search/ alignment are Entrez (NCBI) at http://www.ncbi.nlm.nih.gov/Entrez/ and SRS at http://srs.ebi.ac.uk/

Two important approaches to sequence similarity search/alignment are Profiles (Thompson *et al.*, 1994) and PSI-BLAST (position sensitive iterated-basic linear alignment sequence tool) (Altschul *et al.*, 1997). Profiles express the patterns inherent in a multiple sequence alignment of a set of homologous sequences. The ClustalW algorithm (where W stands for weighting) is one of the most widely used Profiles-based progressive multiple sequence alignment program (Thompson *et al.*, 1994) (ftp://ftp-igbmc.u-strasbg.fr/pub/). This program takes an input set of sequences (the simplest format is fasta format) and calculates a series of pair-wise alignments, comparing each sequence to every other sequence, one at a time. Based on these comparisons, a distance matrix is calculated, reflecting the relatedness of each pair of sequences. The distance matrix, in turn, serves in the calculation of a distant tree (phylogenetic tree) that can be weighted to favor closely related alignment of the most closely related sequences, then realignment with the addition of the next sequence, and so on. The Clustal X (Jeanmougin *et al.*, 1995) is the portable Windows interface Clustal program for multiple sequence alignment.

PSI-BLAST from NCBI is a powerful tool for searching databases for sequences similar to a query sequence. Its earlier program, BLAST goes back to the dotplot approach, checking for well-matching local region. For each entry in the database, it checks for a short contiguous region that matches a short contiguous region of a specific length in the query sequence bidirectionally, using a substitution scoring matrix. Once BLAST identifies a well-fitting region, it tries to extend it. PSI-BLAST begins with a one-at-a-time search to derive pattern information from a multiple sequence alignment of the initial hits, and re-probes the database using the pattern. Then it repeats the process, fine-tuning the pattern in successive cycles as follows:

1. Probe each sequence in the chosen database independently for local regions of similarity to the query sequence, using a BLAST type search but allowing gaps.
2. Collect significant hits and construct a multiple alignment table between the query sequence and the significant local matches.
3. Form a profile from the multiple sequence alignment.
4. Reprobe the database with the profile, continuing to look only for local matches.
5. Decide which hits are statistically significant and retain these significant ones only.
6. Go back to step 2 until a cycle produces no change in the iterated process.

Built on the concept of local alignments, FASTA (Lipman and Pearson, 1985) and BLAST (Altschul *et al.*, 1990) provide increasingly rapid sequence alignments strategies and both can be executed via a personal computer. BLAST uses a heuristic approximation algorithm that gains speed at the expense of accuracy of result. The tool finds short matches between two sequences and attempts to start alignments from these hot spots. There are several types of BLAST with diverse set of features that can add power to the BLAST searching (McGinnis and Madden, 2004). BLAST, which is the main search engine for similarity searching is available at all databases with NCBI BLAST server (http://www.ncbi.nlm.nih.gov/BLAST/) being the most sophisticated with numerous options. All the primary sequence databases provide facilities for multiple alignments of nucleic acids and protein sequences.

14.2.4 Statistical significance of sequence search/alignments

The similarity significance questions whether the similarity has arisen by chance. If the score of the similarity alignment observed is no better than might be expected from a random permutation of the sequence (using a random number generator), then it is likely to have arisen by chance. A score, S can be calculated for any given alignment, but it is important to determine whether this score is high enough to provide evidence of homology. Several statistical indices are often quoted to express the similarity significance in the similarity search and alignment. For pairwise alignments, one of the two sequences is chosen to make many scrambled copies using a random number generator. Each of the permuted (scrambled) copy is aligned to the other unscrambled sequence. For probing a database, the entire database provides a comparison population and the similarity score is given. The commonly used scores are:

- *Raw score and bit score*: The raw score for a local sequence alignment is the sum of the scores of the maximal-scoring segment pairs (MSPs) that make up the alignment. Bit scores are raw scores that have been converted from the log base of the scoring matrix that creates the alignment to ln ($\log_2$)
- *Z-score*: The Z-score is a measure of how unusual the original match is, in terms of the mean and standard deviation of the population scores. If the original alignment has score, S: Z-score of S = (S, mean of population score)/(standard deviation of population). A Z score of zero means that the observed similarity is no better than the average of the control population. The higher the Z scores the greater the probability that the observed alignment has not arisen by chance. As a rule, Z-scores ≥ 5 are significant
- P-*value*: P denotes the probability that the alignment is no better than random. A rough guide to interpreting P values is:

$P \leq 10^{-100}$	exact match
P in the range of 10^{-100}–10^{-50}	sequences very nearly identical
P in the range of 10^{-50}–10^{-10}	homology, closely related sequences
P in the range of 10^{-5}–10^{-1}	usually distant relatives
$P > 10^{-1}$	probably insignificant match

- *E-value*: The E-value of an alignment is the expected number of sequences that give the same Z-score or better if the database is probed with a random sequence. Therefore the E-value depends on the size of the database and varies between 0 and the number of sequences in the database searched. If the statistical significance ascribed to a match is greater than the E-value, the match will not be reported. The lower the E-value, the more stringent the search is (i.e. fewer chance of false positive matches being reported). A rough guide in interpreting E-values is:

$E \leq 0.02$	sequences probably homologous
$E = 0.02 - 1$	homology can not be ruled out
$E > 1$	match by chance

E-values of 0.1 or 0.05 are typically used as cutoffs in sequence database searches.

14.3 MICROARRAY: GENERAL DESCRIPTION [Adapted from Schena 2003]

14.3.1 Introduction

A powerful way to undertake highly parallel analyses of biological samples is to exploit the power of molecular recognition that is to use the specific recognition between two molecules to isolate, detect and identify the target molecule. The most dramatic example of the power of molecular recognition is in the application of DNA microarray (DNA chip) that exploits specific base-pairing between an immobilized target and fluorescently labeled nucleic acid probes to provide rapid simultaneous analysis of thousands of different sequences. A microarray (microscopic array) is an ordered array (collection of uniform sized and spaced analytical elements configured in rows and columns, each with addressable location) of microscopic elements (spots) on a planar substrate that allows the specific binding of genes or gene products (Müller and Nicolau, 2005; Schena, 2000). Thus microarrays (also known as biochips) contain collections of small elements or spots arranged in rows and columns. Molecules in the fluorescent probe react with cognate target sequences on the chip, causing each spot to glow with the intensity proportional to the activity of the expressed genes/gene products. Thus microarrays entail two key components:

1. spatially addressed molecules that recognize individual nucleic acid/protein/glycan probe; and
2. a method to detect the interaction of the individual nucleic acids/proteins/glycans in the mixture with their corresponding recognition molecules.

Miroarray analysis (Schena, 2003) refers to scientific exploration using microarrays. For example, microarrays containing cDNA (typically 500–2000 bp) or oligonucleotides (15–70 nt) as targets produce high specificity and good hybridization signals with fluorescent probe molecules applicable to gene expression profiling and genotyping. Oligonucleotide and cDNA microarrays are nucleic acid microarrays, which encompass microarrays containing any type of DNA or RNA as the target material. Protein microarrays contain pure proteins or cell extracts at each microarray location and glycan (saccharide) microarrays contain saccharides as elements.

Experimental design can be simplified by understanding the five basic steps in the microarray analysis cycle consisting of biological question, sample preparation, biochemical reaction, detection, and data analysis/modeling (Figure 14.3). The e-library (http://arrayit.com/e-library) is an electronic resource and educational tool for microarrays. It contains the full citation, summary information and a link to the electronic manuscript. Links to electronic manuscripts allow the user to navigate to sites for further information.

14.3.2 Surface preparation for microarray

A successful sample preparation ensures robust microarray analysis, which requires high quality surfaces for the reaction between target and probe. Of the many different types of

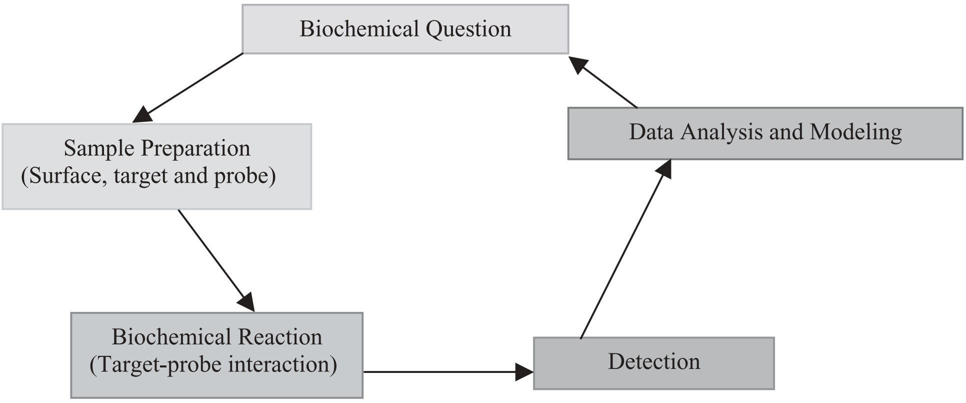

Biochemical Question
Sample Preparation
(Surface, target and probe)
Biochemical Reaction
(Target-probe interaction)
Detection
Data Analysis and Modeling

Figure 14.3 Microarray analysis cycle

surfaces used for microarray analysis, the most common treatment involves an introduction of chemically reactive amine or aldehyde functionalities, which allow attachment of biomolecules of interest by electrostatic interactions or covalent bonds respectively. The reaction of glass with 3-aminopropyltrimethoxysilane provides the amine surface (Figure 14.4). Hydroxyl groups present on the glass surface act as nucleophiles that attack the silicon atoms, resulting in covalent bond formation creating the surface that contains reactive amine groups. Some organosilane compounds that can be employed are: aminopropylsilanetriol, 3-aminopropylmethyldiethoxysilan, allylaminotrimethylsilane, 3-(2-aminoethylamino)propyltrimethoxysilane, 4-aminobutyltriethoxysilane, 3-aminopropylpentamethyldisiloxane, (aminoethylaminomethyl)phenethyltrimethoxysilane, *N*-(2-aminoethyl)-3-aminopropylmethyldimethoxysilane, 3-aminopropyltrimethoxysilane, *N*-(6-aminohexyl)aminopropyltrimethoxysilane, aminomethyltrimethylsilane and 3-(*m*-aminophenoxy)propyltrimethoxysilane. The overall charge of amine microarray surfaces allows attachment of printed biomolecules that carry negative charges via electrostatic interactions. Attachment of nucleic acids to an amine surface occurs by interactions between negatively charged phosphate groups on the DNA backbone and positively charged surface amine groups termed adsorption (Figure 14.4A). The electrostatic interactions occur between the surface amines and the carboxyl groups of C-terminus, Asp and Glu of proteins. Amine surfaces tend to perform well over a wide range of experimental conditions and generally manifest both low intrinsic fluorescence (before hybridization) and low background fluorescence (after hybridization).

Aldehyde surface preparation uses silane reagent to form covalent bonds between silicon atoms on the microarray substrate and organic reactive groups. An aldehyde surface typically contains a spacer arm of approximately 10 carbon atoms at the end of which is a reactive aldehyde group. The DNA target must be derivatized first with a primary amine attached via a 6–12 spacer arm to the phosphate group of the 5′-nucleotide of the DNA molecule. The N-terminal amino and ε-amino groups of proteins provide the reactive sites for the linkages. Subsequent treatment with $NaBH_4$ reduces the Schiff base to secondary amine and unreacted aldehyde groups to alcohol counterpart (reducing background fluorescence). The attachment to an aldehyde surface involves covalent bond formation; therefore the binding to this surface is termed covalent coupling. Table 14.6 compares the two types of microarray surfaces.

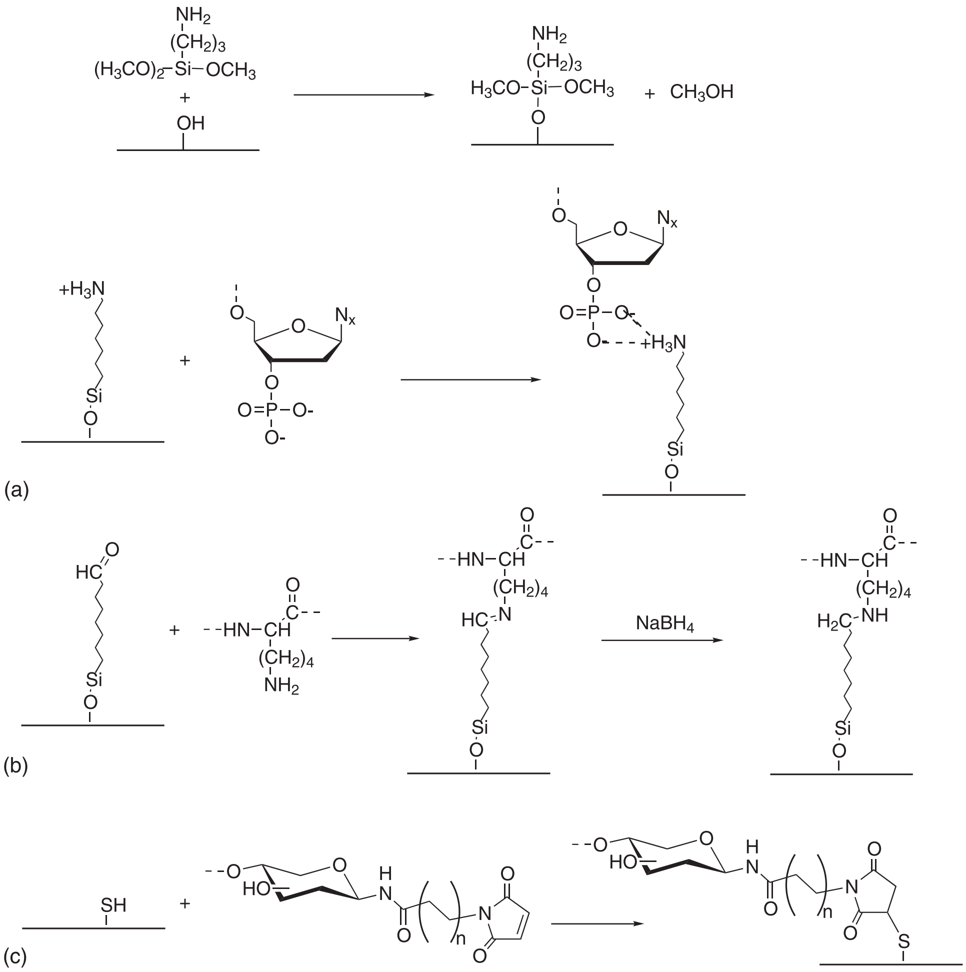


Figure 14.4 Attachment of targets to microarray
The attachments of target molecules to the microarray surfaces are exemplified for ionic interaction between 5′-phosphate anions of oligonucleotide and amino cation (A), covalent formation between ε-amino group of lysine residue of oligopeptide and aldehyde group *via* Schiff base formation followed by reduction (B) and hetero-Michael addition between maleimide conjugate of mono-/oligosaccharides and thiol group (C).

TABLE 14.6 Comparison of amine and aldehyde surfaces for biochips

Criterion	Amine surface	Aldehyde surface
Surface charge	Positive	Neutral
Intrinsic fluorescence	Very low	Very low
Surface character	Slightly hydrophilic	Hydrophobic
Printed sample spreading	Slight	Slight
Linkers requirement	Azide for glycans	Primary amine for DNA
Attachment mechanism	Adsorption, electrostatic	Covalent
Attachment chemistry	Electrostatic for DNA and proteins, covalent for glycans	Covalent *via* Schiff base, reduction
Attachment geometry	Nonspecific/specific for glycans	Specific attachment
Assay sensitivity	Excellent	Excellent
Assay background fluorescence	Very low	Low

Taken from Schena (2003)

Note: The characteristics of expoxy surface is analogous to aldehyde surface.

14.3.3 Microarray targets

Target is defined as the molecule tethered to a microarray substrate that reacts with a complementary probe molecule in solution. There are two approaches to microarray target preparation, known as delivery and synthesis. The delivery approaches prepare the target molecules off-line (e.g. DNA by PCR, proteins from monoclonal antibodies). Subsequently the target molecules are delivered on to the microarray surface by contact printing or an equivalent means. The synthetic methods create target elements directly on the microarray surface by joining monomer building blocks together via a solid phase (subsection 8.4) or combinatorial strategy (subsection 8.5). The delivery approaches have several advantages including ease of implementation, availability/affordability, macromolecular diversity and quality, whereas the synthesis approaches to target preparation are favored in several ways including high density, cost effectiveness at high target complexity and reliable target identification.

For DNA microarrays, the PCR process provides microgram quantities of target material for any DNA segment of interest from any organism (Innis *et al.*, 1999). PCR products for microarrays are generated using either common primers or gene-specific primers and are attached to the functionalized surface (Figure 14.1A). In common primers, two oligonucleotides used for PCR amplification bind to a vector sequence and allow amplification of all of the targets using a single oligonucleotide pair. This allows the amplification of cDNAs or expressed sequence tags (ESTs) from any available library to create microarrays of gene segments from any tissue or organism. Gene-specific primers are used to obtain unique targets from the genomic DNA. Two specific oligonucleotide pair of primers bind uniquely to a given target sequence and then allow amplification of a unique target from a complex mixture of complementary or genomic DNA. The PCR products are relatively large size (500–5000 bp) that provides extensive complementarity for hybridization, generating intense fluorescent signals. However, the PCR products are double-stranded nucleic acid targets that require denaturation (typically with boiling water for few minutes) before use. The presence of complementary target strands in the PCR products means that the re-annealing or snapping back can occur during the course of the hybridization. Single-stranded synthetic oligonucleotides provide another common source of target sequences for nucleic acid microarrays. They can be prepared using either delivery (5–120 nucleotides) or synthesis (5–25 nucleotides) strategies with conventional phosphoramidite synthesis (Caruthers, 1985) or one of the *in situ* approaches such as solid-phase chemical synthesis (subsection 8.4) and photolithography that used ultraviolet light (subsections 8.5). The synthetic oligonucleotides allow unique targets to be made for large gene families, highly homologous sequences, and mRNA from the same gene that differ solely on the basis of mRNA splicing or processing.

Protein microarrays are prepared from diverse sources. The N-terminal and side chain of Lys are primary amines that can attach to microarray substrates by Schiff base interactions with reactive aldehyde groups followed by $NaBH_4$ reduction to form covalent linkages (Figure 14.4B). The negatively charged C-terminal, Glu and Asp residues bind to microarray substrates via ionic interactions with reactive amine surfaces. The concentration of the printed targets is 10 mg/mL. The most popular target molecules for protein microarray are monoclonal antibodies or recombinant IgG fragments, which are prepared by the use of phage display libraries (Vaughan *et al.*, 1996). An alternative approach involves immobilization of proteins (e.g. IgG) within polyacrylamide gel pads arrayed on silanized glass. The grid is generated by masking with a nontransparent chromium film and the gel matrix is activated by treatment with glutaldehyde and a solution of IgG (~1 nL) on to each gel pad (Guschin *et al.*, 1997).

For a carbohydrate microarray, a glass slide is modified by the thiol group as solid support. Carbohydrates in the form of glycosylamines are converted into maleimide via a hydrocarbon tether chain and covalently bound to the glass surface by hetero-Michael addition reaction (Park and Shin, 2002) between the thiol group and the maleimide moiety of the glycosyl derivative (Figure 14.4C).

Molecular imprinting involves the synthesis of artificial recognition sites on a surface by mimicking the shape of the template molecule in a polymeric film, thereby forming a molecularly imprinted polymer (MIP). It is assumed that any biomacromolecular interactions in which shape plays a part, such as biomacromolecule-ligand bindings, may be mimicked by MIPs (Haupt and Mosbach, 1998). The template molecule with the polymer reagents allows the matrix to harden and the template is removed with a specific solvent. This leaves holes in the polymer that mimic the shape of the template molecule and may be employed to capture the template-shaped target molecules from a mixture. The MIP library can be constructed by employing a combinatorial library of templates (Takeuchi *et al*., 1999) and used to prepare protein/glycan microarrays, whichever steric interaction may be involved in the molecular recognition.

Microarray manufacturing technology that uses miniaturization strategy from the computer chip industry, including photolithography, is known as semiconductor technology. All the semiconductor approaches use solid-phase synthesis to build the microarray on the chip in a stepwise manner using the four nucleotides/twenty amino acids. Because chromium masks and nucleotides/amino acids can be combined in many different ways to achieve extensive synthetic diversity, this semiconductor approach is commonly known as the combinatorial method as described earlier (subsection 8.5). Photolithography uses ultraviolet light and solid-phase chemical synthesis to manufacture microarrays. Photolithography of peptide synthesis has been described. An analogous approach can be employed to synthesize oligonucleotides/oligopeptides for DNA/protein microarrays.

In an alternative approach, photodeposition chemistry based on maskless array synthesizer (MAS) uses a maskless light projector as a virtual mask to fabricate microarrays. The virtual mask is an array of hundreds of thousands of individually addressable aluminum mirrors on a computer chip. These mirrors function as virtual masks that reflect the desired pattern of UV light and are controlled by the computer.

14.3.4 Microarray probes

A probe is defined as a labeled molecule in solution that reacts with a complementary target molecule on the substrate. The maximum probe concentration represents the strongest signal attainable within the psudo-first region of a kinetic plot (signals versus probe concentrations). Fluorescence is the dominant probe label for microarray analysis. Probes can be prepared by either direct labeling or indirect labeling. The direct labeling schemes attach fluorescent tags (section 7.3) in a covalent manner to the probe molecules by enzymatic or chemical reactions, whereas the indirect labeling approaches attach fluorescent tags in a noncovalent manner to the probe molecules via a bridge molecule which, in turn, is attached to a fluorescent reagent. Probes for DNA microarrays are fluorescent labeled by the polymerase or reverse transcriptase catalyzed reaction or its modified procedure. Polymerase/reverse transcriptase catalyzes the synthesis of DNA/mRNA/cDNA in the presence of fluorescent nucleotide or aminoallyl nucleotide analogs. This incorporates the fluorescent nucleotide into DNA/mRNA/cDNA or nucleotide analogs, which can be coupled with fluorescent dyes.

For proteins microarrays, most protocols link fluorescent dyes to proteins in a covalent bond via Lys and Arg residues. Fluorescein isothiocyanate (FITC) contains a reactive isothiocyanate group that is susceptible to nucleophilic attack by primary amines on the surface of proteins, resulting in covalent attachment of the fluorescent dye. Because Lys and Arg residues make up nearly 10% of the amino acids in a given protein, purified proteins and cellular extract can be labeled with high efficiency with FITC. A labeling reaction containing 50 μg FITC and 1.0 mg protein usually provides brightly fluorescent protein mixtures. Cyanine (Cy) dyes are available as amine-reactive derivatives, allowing the cyanine dyes to be conjugated to proteins via Lys and Arg residues. The fluorescent protein, phycoerythrin (PE) is widely used in microarray probe labeling to provide an intense florescence. Phycoerythrin is first reacted with maleimide-PE such that the free sulfhydryl groups on the protein act as nucleophiles, attacking the maleimide group on the PE and coupling fluorescent PE to the protein probe.

14.3.5 Biochemical reaction of microarray

The specificity and affinity of interactions between target molecules bound to the microarray substrate and probe molecules in solution largely determine the quality of microarray assays. The complementary base hybridization is the most efficient and reproducible target-probe interactions used in DNA microarray analysis.

Sequence composition is important in hybridization reactions involving short targets and probes (e.g. oligonucleotides). Hybridization affinity and signal intensity correlate directly in microarray assays, with the G:C sequences producing extremely intense fluorescence and the A : T(U) sequences producing much weaker fluorescence. Sequence composition is a minor consideration if the targets and probes are long (e.g. cDNA). Long heteroduplexes have a much greater affinity than short heteroduplexes.

Microarray analysis provides a technology platform for massive, parallel analysis of protein–protein interactions (MacBeath and Schrieber, 2000; Joos *et al*., 2002). The difficulty with protein microarray is that proteins do not behave as uniformly as nucleic acids. Protein function is dependent on a precise and fragile 3D structure that may be difficult to maintain in a microarray format. A practical challenge posed by proteins derives from their relatively delicate tertiary structure, which is susceptible to unfolding during microarray printing. In addition, the efficiency and specificity of protein–protein interactions are not nearly as standardized as nucleic acid hybridization. The use of lectin microarrays in glycoform analysis also encounters the similar problems.

14.3.6 Microarray detection

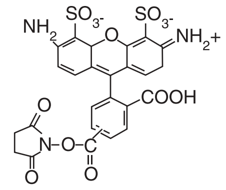
The use of fluorescent detection scheme (section 7.3) allows multiple biological samples to be examined on a single microarray and provides a high level of detection. Microarray assays use dyes with large Stokes shifts and high fluorescent intensity. The derivatives of FITC containing polar, alkyl and heterocyclic groups with distinct absorption and emission spectra, which enable two or more labels to be used simultaneously, have been the most widely used dyes in microarray analysis (Table 14.7). The cyanine molecules including indocarbocyanine and indodicarbocyanine, which demonstrate large molar extension coefficients and convenient emissions are also widely used fluorescent dye family. They are related to the commercial cyanine dyes, Cy3 and Cy5, which contain two sulfate groups and therefore are soluble in aqueous solutions. Cy3 and Cy5 are also available as phosphoramidite derivatives and can thus be incorporated into oligonucleotides and used directly as hybridization probes in microarray experiments. The BODIPY series form another dye family used in microarray analysis.

TABLE 14.7 Fluorescent dyes used in microarray detection

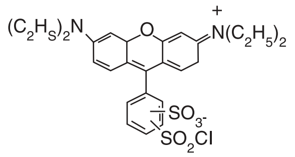
Dye	Absorption	Emission	Quantum yield	Coefficient
DAPI	350	450	0.83	120 000
FITC	490	520	0.71	150 000
Alexa 488	495	519		
Rhodamine 6G	525	555	0.9	85 000
B-phycoerythrin	546	575	0.98	2 410 000
R-phycoerythrin	546	578	0.98	1 960 000
Cy3	550	580	0.14	150 000
Lissamine	570	590		
Alexa 568	578	603		
Texas Red	596	620	0.51	85 000
BODIPY 630/650	625	640		
Cy5	649	670	0.15	250 000
Cy7	743	767		250 000

Taken from Schena (2003)

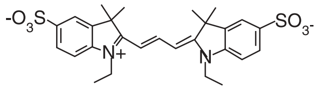
Note: Structures of some listed fluorescent dyes are:



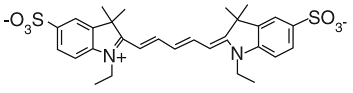
Alexa 532



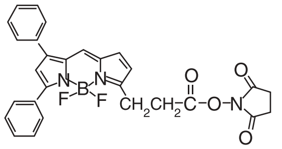
Lissamine



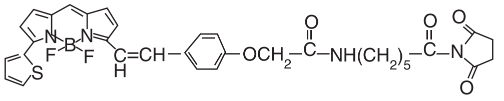
Cy3



Cy5



BODIPY 575/589



BODIPY 630/650

14.3.7 Data analysis in microarray

Microarray scanners and imagers acquire fluorescent microarray data as tagged image file format (TIFF) files. The TIFF file (.tif) is a two-dimensional intensity map of the microarray surface, with fluorescent signals stored as pixels. The process by which numerical values are obtained from flat microarray data files is known as quantitation or quantificaiton. The numerical values provide information that can be used to determine the concentration of genomic DNA, mRNA, protein and other biomolecules of interest in complex samples. The process by which microarray signals are demarcated from background in a microarray image is known as signal segmentation. The user can increase the precision of signal segmentation by adjusting the physical and statistical parameters used to demarcate signal and background. The process by which signal intensities from two microarray images are divided to obtain a quotient for each datum point is known as a ratio calculation. It is used widely in microarray analysis to obtain quantitative information concerning genotypes, gene/protein expression patterns and other biological processes (Stekel, 2003).

Normalization is the process by which data from different channels or different chips are equalized before analysis and the value that is used to normalize different datasets is known as a normalization factor. Normalization corrects minor imbalance that arises during the imaging process, and can be accomplished using a variety of different criteria, including global intensities, housekeeping genes and internal standards. Global intensity normalization uses the sum of signals in multiple images to provide equalized signals. Normalization of data from different biological tissues or from many different microarray experiments, uses a set of cellular genes known as housekeeping genes (a housekeeping gene refers to the cellular gene that plays a central role in all cells and correspondingly is expressed nearly equally in all cells or tissues). Another approach is the use of a standard curve to assign quantitative values to genes or gene products of unknown concentration.

Data obtained from microarray detection instruments are expressed in units of fluorescent counts with 16-bit values ranging from 1 to 65 536. It is common practice to convert or transform the raw counts into a logarithmic scale. Microarray data transformed into a log scale (0–4.8) exhibit a more uniform distribution than raw signal intensities (1–65 536).

One of the most useful representations of biomacromolecular expression data is the scatter plot. A scatter plot is a two-dimensional representation of microarray data in which the signal intensities of two samples are plotted. Scatter plots facilitate visualization of two samples in two-dimensional space, and a three-dimensional scatter plot could be used to compare three samples. For example, genes expressed at a high level (abundant mRNAs) reside at a greater distance from the origin than genes expressed at a low level (rare mRNAs) in the scatter plot. Genes that are activated or repressed fall above and below the diagonal or identity line respectively.

For comparing expression profiles in many samples, techniques that allow dimensionality reduction to facilitate graphical visualization of many comparisons are employed. Reducing complex comparisons into three-dimensional space allows representation of the data from hundreds of different samples on a conventional computer screen. One of the most common methods of dimensionality reduction is principle component analysis (PCA). PCA is a method that reduces relationships that exist in high-dimensional space into three dimensions, enabling complex data to be visualized in standard graphical form. PCA is known as a multivariate technique, because it is a statistical approach that allows the comparison of many different variables (genes/proteins/glycoforms). The PCA method preserves closeness relationships between variables in multidimensional space, so that variables residing in close proximity in many dimensions are configured close to each other in three dimensions (Kim *et al*., 2001). Another multivariate classification method is known as cluster analysis, which is a general means of categorizing data based on the similarity of the datum points to one another. Clustering places the similar data into a cluster and then organizes the cluster with respect to one another, so that similar clusters are close to one another and dissimilar groups are far apart (Eisen *et al*., 1998).

Cluster analysis becomes an increasingly important tool for gene/protein expression analysis in microarrays (Eisen *et al*., 1998; Kaufman and Rousseeuw, 1990). Genes/proteins that share the most common pattern of expression are placed into a single cluster. The tight correlation between expression and function means that genes/proteins that cluster together often share a common function. Clustering therefore provides a rapid and intuitive means of identifying and visualizing genes/proteins that share a similar function.

Clustering also provides a way to ascribe putative function to novel sequences if those sequences fall into a cluster that contains genes/proteins of well-known functions. Thus clustering provides a rapid means of establishing a putative function for genes/proteins that have no known function for which functional roles have yet to be established. Changes in gene/protein expression embody both a quantity (mRNA concentration/bioactivity) and a direction (up or down), making a mathematical vector a well-suited approximation for their expressions. A vector is a mathematical quantity that has both a magnitude (e.g. fold change) and a direction (e.g. activation or repression) and is thus a good descriptor for gene/protein expression.

There are two broad categories of clustering. Supervised clustering exploits known reference vectors to classify and organize expression data, whereas unsupervised techniques do not use reference vectors. Clustering methods are also classified as hierarchical and nonhierarchical. Hierarchical clustering methods establish a small group of genes/proteins that share a common pattern of expression and then construct the dendrogram in a sequential manner using a ranked series of hierarchy of cluster. In nonhierarchical methods, such as k-means clustering, each expression datum is assigned to a cluster based on its expression profile, and the process is repeated until every datum point has been placed into a cluster. The user initiates the k-means process by specifying the number of cluster (k) into which the genes/proteins are to be assigned, and the computer constructs a dendrogram of the clusters by iterative processes. In addition to viewing clustered information in dendrogram and cluster formats, it is possible to display cluster information in a two-dimensional manner using self-organizing map (SOM) algorithms. The SOM is a visualization tool that displays each cluster in a separate software window, allowing the user to specify the physical layout.

14.4 COMPUTER TECHNOLOGY

14.4.1 Machine: Computer

The biochemists are most likely to be working with one of the three types of computers: personal computer (PC), notebook or handheld computer (Becker *et al*., 2001; Tsai, 2002), although workstations with improved capacity, computing speed and other added facilities are used in dedicated research laboratories and networked internally to form a local area network (LAN). A computer system (Mueller, 2005; Meyers and Jernigan, 2004; Weber, 2004) consists of internal and external devices. Inside a computer system box (chassis) is:

- Motherboard: It is a printed circuit board that is the main circuit board of a computer. The motherboard houses:
    - central processor unit (CPU), which executes computer instructions;
    - random access memory (RAM) or the main memory, which is used to store data and instructions;
    - basic input/output system (BIOS), which initializes and managing the basic input and output devices;
    - buses or traces which are paths to carry data for the processor, BIOS and the memory;
    - controllers, which control the operation of peripheral devices. Controllers are built as single chips. All the chips on the main board are collectively known as the chip set.

  - connector, which is a part of a cable that plugs into a port or an interface connecting one device to another;
  - ports, which are external connectors for external devices. Different types of ports are PS2 port, parallel port, serial port and universal serial bus (USB) ports;
  - expansion bus slots for insertion of an adapter board or card.
- Internal disk drive
- Adapter or expansion boards
- Power supply
- Fan

The external devices of a computer are those that can be used without opening the system's box. They are:

- input device such as keyboard, pointing device (mouse, track ball), tablet;
- display device such monitor screen, LCD (liquid crystal display),
- other peripherals such as printer, scanner, camera.

The CPU is the brain of the computer, the input and output devices are its sensors to the outside world and the memory devices hold the information to be processed. These components work together as schematically represented in Figure 14.5:

The computer gets input from the keyboard, or a tablet in the form of input characters. A pointing device such as the mouse or a track ball permits menu selections. Most of the calculation and execution of program instructions take place in the CPU, which is also responsible for storing and retrieving information on disks and other media. Within CPU, the Arithmetic Logic Unit (ALU) performs arithmetic and logic operations. The arithmetic operations include addition, subtraction, multiplication and division. The logic operation is the comparison operation for >, =, <, ≤ or ≥. The Control Unit (CU) of the CPU communicates with the memory and the ALU to direct the execution of program instructions. It fetches instructions from the main memory, decodes and executes them. However, if an instruction requires calculations or a comparison, it sends a signal to call on the ALU to perform the operation. Thus the CU is responsible for controlling and synchronizing the action of the CPU. Internal Registers is the high speed internal memory

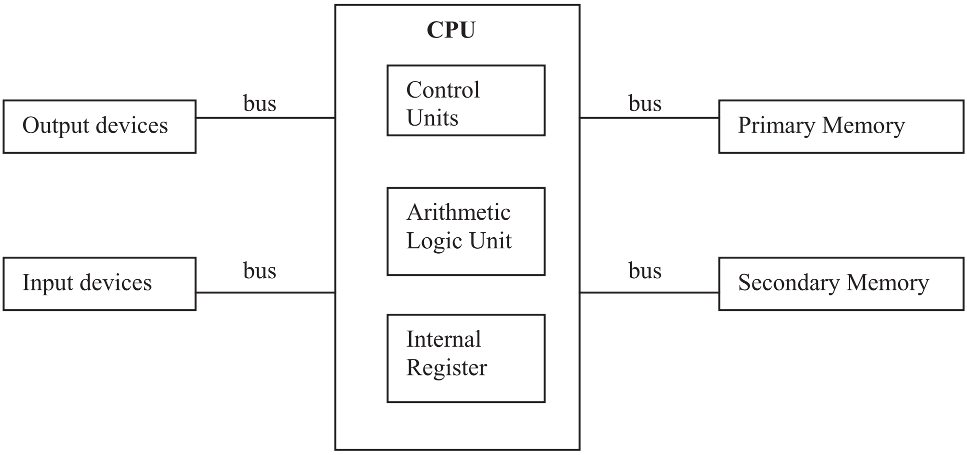


Figure 14.5 Architecture of a computer system

within the CPU and is used as temporary storage area for instructions or data. Register work under the direction of the CU to store data and instruction. Each of the following different types of registers has a special role:

- ❏ Accumulation stores the intermediate result of computation;
- ❏ Memory Address Register (MAR) is used to hold the address of an instruction or a piece of data that are transferred from RAM;
- ❏ Memory Buffer Register (MBR) temporarily holds data taken/send from/to the main memory;
- ❏ Instruction Register (IR) contains the instruction to be executed by the CPU;
- ❏ Program Counter tracks the location of the next instruction to be executed in the program.

Figure 14.6 depicts the steps and components that are involved in executing a program instruction by the CPU. An instruction cycle is the time period during which an instruction is fetched from the main memory and executed by the CPU. The CPU carries out an execution of an instruction in four steps:

1. *Fetch cycle*: The address of the instruction is placed in the MAR to fetch it from RAM into the MBR and then the IR when the CU is ready;
2. *Decoding*: The CU decodes the instruction in IR into an op code and operands;
3. *Fetch cycle*: If an operand involves a memory access, fetch the operand from the main memory by loading the address of the operand into the MAR; and
4. *Execution*: The ALU execute the instruction when the operands are ready.

14.4.2 Tool: Program, language and programming

The Operating System (OS) is a collection of software programs that control the operations of a computer system. The OS acts as a necessary software layer between the computer's hardware and the programs, by managing requests and communication between them. It gives the intelligence to the computer. The dominant OS for personal computer (PC) is Microsoft's Windows (Meyers and Jernigan, 2004; Weber, 2004). The Linux Operating System has been making inroads into the PC. The functions of the OS programs include:

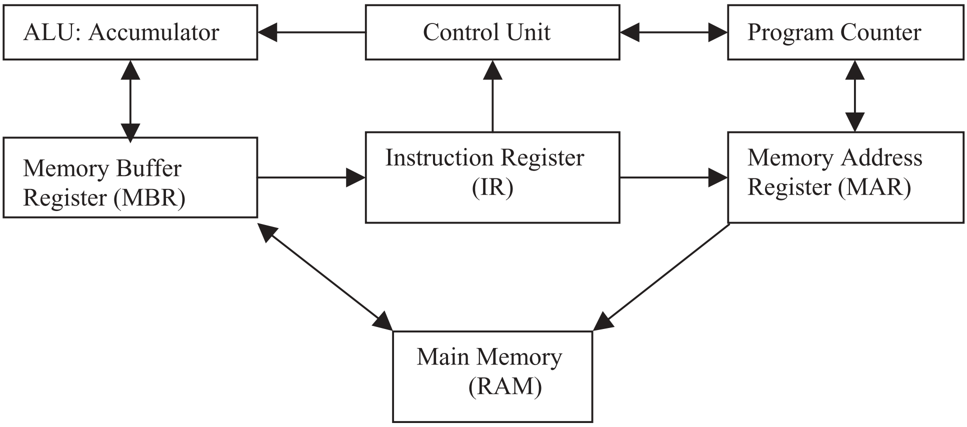

ALU: Accumulator
Control Unit
Program Counter
Memory Buffer Register (MBR)
Instruction Register (IR)
Memory Address Register (MAR)
Main Memory (RAM)

Figure 14.6 Instruction execution cycle

1. *Manage computer resources*: Computer resources such as the main memory, input/output devices, secondary storage devices, and the processor's time are managed.
2. *Provide command languages*: The OS provides command languages such as DOS commands for Microsoft OS or shells for Unix or Linux. Window provides graphical user interface (GUI) that allows the user to issue commands to the computer by point-and-click.
3. *Provide networking facilities*: PC can be configured to connect to the Internet using the communication protocol, TCP/IP provided the user has an account with an Internet Service Provider (ISP) that channels the computer connection into the Internet highway.
4. *Provide software for program development*: The OS provides the following software for the development and running of the program:
    - a text editor e.g. notepad, textpad, nedit, etc.;
    - an interpreter or a compiler conferring a program code into the machine language of the computer;
    - a linker linking different modules of a program and library functions;
    - a loader loading an executable file into the main memory;
    - a scheduler dispatching a program for the ready-to-run queue.
5. *Utilities*: Some utilities that come with the OS are system tool, communications, editors and multimedia players.

The program or instructions for the computer must be written in the language that is executable by the computer. The bioinformatics programs are generally written in high level languages such as C, C_{++}, JAVA and FORTRAN. The choice of programming language depends on the nature of the algorithm and associated data structure, and the expected use of the program. The primary objective for most biochemists is developing proficiency in using the operating system of one's computer and tools available on the Web to solve biochemical problems. For advanced students, some knowledge in Practical Extraction and Report Language (Perl) may be useful in order to sift through tons of data and to extract the required information (Jamison, 2003). Perl, with its capacity to detect patterns of data, especially strings of text, has a variety of attributes (such as descriptive and relatively easy to learn, smooth integration on to Unix-base systems and cross-platform portability) that make it popular in bioinformatics. Perl can be downloaded from http://www.bioperl.org. A Perl tutorial can be accessed from the same site, which also provides links to many module documents (ready to use programs).

Any sequence of computational steps performed by a computer is known as an algorithm. An algorithm that provides an approximate solution to an extremely complex problem is known as a heuristic algorithm (e.g. BLAST). Heuristics can be applied in cases where correct solutions are impractical to obtain by computer. For the sequence alignment, the Needleman–Wunsch (1970) and Smith–Waterman (1981) algorithms employ dynamic programming to find optimal sequence alignments. A dynamic programming algorithm finds the best solution by first breaking the original problem into smaller subprograms and then solving. Dynamic programming works, by first solving all these subproblems, storing each intermediate solution in a table along with a score, and finally choosing the sequence of solutions that yields the highest score. For sequence alignment, it essentially tackles the alignment problem backward and finds the best alignment by making decisions with a series of sub-alignments scored one after another. Needle-

TABLE 14.8 Useful biocomputing freeware for Windows

Software program	URL	Features
OpenOffice	http://www.openoffice.org/	Office suite: WP, SS, DB, editing formula and drawing
ViSta	http://forrest.psych.unc.edu/research/	Statistical analysis package
GIMP	http://gimp-win.sourceforge.net/	Photo-editing, image manipulating program
Ghostscript	http://www.gostscript.com/doc/AFPL	Conversion of given file format into the PDF via PS
Gepasi	http://www.gepasi.org	BD: kinetics, metabolic modeling and simulation
SimFit	http://www.simfit.man.ac.uk/default.htm	BD: kinetics, simulation, statistical analysis, curve fit and graph plot
BioEdit	http://www.mbio.ncsu.edu/BioEdit/bioedit.html	Sequence alignment editor
SeWeR	http://www.bioinformatics.org/sewer	Sequence analysis package
Swiss-Pdb Viewer	http://www.expasy.org/spdbv/text/download.htm	Biomacromolecular modeling
VMD	http://www.ks.uiuc.edu/Development/Download/	Molecular modeling and informatics
Biodesigner	http://www.prix.com/biodesigner.download.html	Molecular modeling

Note: Abbreviations used: BD, biochemical dynamics; DB, database; PDF, portable document format; PS, postscript; SS, spread sheets, WP, word processing.

man–Wunsch uses a global alignment strategy in which the entire sequences of the genes or proteins are compared, whereas Smith–Waterman employs local alignment method, allowing for comparisons and alignments based on small regions of sequence similarity.

With the availability of reasonably priced high-power PC and advances in both accessibility and maturation of freeware and open-source software, it is now possible to create highly capable computer systems that enable biochemists to set up a complete working biocomputing platform. Open-source software is being developed with the programming code freely available for anyone to obtain, scrutinize and use (see the Open-Source Initiative at http://www.opensource.org), while freeware is usually closed-source software that is being distributed for free (see the Free Software Foundation at http://www.fsf.org). Table 14.8 lists some useful freeware programs for biocomputing.

14.4.3 Molecular graphics

Molecular graphics (Henkel and Clarke, 1985) refers to a technique for the visualization and manipulation of molecules on a graphical display device. The technique provides an exciting opportunity to augment the traditional description of chemical structures by allowing the manipulation and observation in real time and in three dimensions, of both molecular structures and many of their calculated properties. Recent advances in this area allow visualization of even intimate mechanisms of chemical reactions by graphical representation of the distribution and redistribution of electron density in atoms and molecules along the reaction pathway.

All graphics programs must be able to import commands defining representations and translate these into a picture according to the representations specified. The graphics programs offer various choices of renderings of the model, with color coding of atoms or groups, and with selective labeling. These include the followings:

14.4.3.1 Line drawings: skeletal and ball-and-stick models. Traditionally, drawings of small molecules have represented either (a) each atom by a sphere or (b) each bond by a line segments. Bond representations give a clearer picture of the topology or connectivity of a structure. In a simple picture of line drawings, there is a direct correspondence, i.e. one line segment equals one bond. A ball-and-stick drawing is a simple skeletal model, in which the representation of the bond is generalized to a cylinder and a disc is added at the position of each atom. Additional information may thereby be displayed in that different atom types may be distinguished by size and shading, and bonds of different appearance are drawn. Wire models and ball-and-stick models are extremely useful because of the great detail they contain. They are particularly useful in connection with blow-ups of selected biomacromolecules.

Each atom is shown to represent complete chemical detail of molecules in shaded-sphere pictures. This category includes three basic types of pictures:

1. *Line drawings*: Each atom is represented as a disc. The picture is a limit of the ball-and-stick drawings as the radius of each atomic ball is made in proportional to the van der Waals radius.
2. *Color raster devices*: A raster device can map an array stored in memory on to the screen so that the value of each element of the array controls the appearance of the corresponding point on the screen. It is possible to draw each atom as a shaded sphere, or to simulate the appearance of the Corey–Pauling–Koltun (CPK) physical models to maintain most of the familiar color scheme (i.e. C = black, N = blue, O = red, P = Green, and S = yellow). In such representation, atoms are usually opaque, so that only the front layer of atoms is visible. However, clipping with an inner plane or rotation can show the packing in the molecular interior.
3. *Real-time rotation and clipping*: Facilities available on vector graphics devices are useful in connection with another technique for representing atomic and molecular surfaces. The spatter-painting of the surface of a sphere by a distribution of several hundred dots produces a translucent representation of the surface (dot-surface). It is possible to combine dot-surface representations with skeletal models to show both the topology of the molecule and its space-filling properties. Dot-surface pictures used on an interactive graphics device with a color screen, have been helpful in solving problems of docking ligands to proteins and exploring the goodness of fit in interfaces.

There are three levels of structural graphics; 1D formula or string/character format (e.g. SMILES string), 2D chemical structures (e.g. ISIS draw) and 3D molecular structures (e.g. PDB atomic coordinate display). The general rules for representing chemical and biomonomer structures by SMILES (Weininger, 1988), which is recognizable by a number of 2D drawing programs, can be accessed at http://www.daylight.com/dayhtml/smiles/. For biomacromolecules, the primary structures represented in coding sequences are 1D format. The common 1D presentations for nucleotide sequences are in GenBank, EMBL/EBI and fasta formats (Figure 14.7). The common files for amino acid sequences are PIR, Swiss-Prot (EMBL/EBI), GenPept and FASTA formats (Figure 14.8). ReadSeq (http://dot.imgen.bcm.tmc.edu:9331/seq-util/Options/readseq.html) provides a facility for interconverting different sequence formats. The linear code (Banin *et al*., 2002) for glycans is described in subsection 6.1.2.

The sequences represented in the chemical linkages of nucleotide/amino acid/glycose structures are 2D drawings. The nucleotide/amino acid sequences in character format (without index, e.g. fasta format) can be converted into the 2D structures with ISIS Draw (Figure 14.9), which can be downloaded from MDL Information System at http://www.mdli.com/download/isisdraw.html

```
BASE COUNT      132 a    169 c    172 g    111 t
ORIGIN
        1 tcccgctgtg tgtacgacac tggcaacatg aggtctttgc taatcttggt gctttgcttc
       61 ctgcccctgg ctgctctggg gaaagtcttt ggacgatgtg agctggcagc ggctatgaag
      121 cgtcacggac ttgataacta tcggggatac agcctgggaa actgggtgtg tgttgcaaaa
      181 ttcgagagta acttcaacac ccaggctaca aaccgtaaca ccgatgggag taccgactac
      241 ggaatcctac agatcaacag ccgctggtgg tgcaacgatg gcaggacccc aggctccagg
      301 aacctgtgca acatcccgtg ctcagccctg ctgagctcag acataacagc gagcgtgaac
      361 tgcgcgaaga agatcgtcag cgatggaaac ggcatgagcg cgtgggtcgc ctggcgcaac
      421 cgctgcaagg gtaccgacgt ccaggcgtgg atcagaggct gccggctgtg aggagctgcc
      481 gcacccggcc cgcccgctgc acagccggcc gctttgcgag cgcgacgcta cccgcttggc
      541 agttttaaac gcatccctca ttaaaacgac tatacgcaaa cgcc
//

SQ    Sequence 584 BP; 132 A; 169 C; 172 G; 111 T; 0 other;
      tcccgctgtg tgtacgacac tggcaacatg aggtctttgc taatcttggt gctttgcttc  60
      ctgcccctgg ctgctctggg gaaagtcttt ggacgatgtg agctggcagc ggctatgaag 120
      cgtcacggac ttgataacta tcggggatac agcctgggaa actgggtgtg tgttgcaaaa 180
      ttcgagagta acttcaacac ccaggctaca aaccgtaaca ccgatgggag taccgactac 240
      ggaatcctac agatcaacag ccgctggtgg tgcaacgatg gcaggacccc aggctccagg 300
      aacctgtgca acatcccgtg ctcagccctg ctgagctcag acataacagc gagcgtgaac 360
      tgcgcgaaga agatcgtcag cgatggaaac ggcatgagcg cgtgggtcgc ctggcgcaac 420
      cgctgcaagg gtaccgacgt ccaggcgtgg atcagaggct gccggctgtg aggagctgcc 480
      gcacccggcc cgcccgctgc acagccggcc gctttgcgag cgcgacgcta cccgcttggc 540
      agttttaaac gcatccctca ttaaaacgac tatacgcaaa cgcc                 584
//
```

Figure 14.7 Database file formats for nucleotide sequence
GenBank (upper) and EMBL (lower) formats for nucleotide sequence of chicken lysozyme gene are represented. The end of the file is indicated by //. Fasta format is shown in Subsection 4.1.2.

```
SUMMARY         #length 147 #molecular_weight 16238
SEQUENCE
            5          10        15        20        25        30
    1 M R S L L I L V L C F L P L A A L G K V F G R C E L A A A M
   31 K R H G L D N Y R G Y S L G N W V C A A K F E S N F N T Q A
   61 T N R N T D G S T D Y G I L Q I N S R W W C N D G R T P G S
   91 R N L C N I P C S A L L S S D I T A S V N C A K K I V S D G
  121 N G M N A W V A W R N R C K G T D V Q A W I R G C R L


Length: 147 AA    Molecular weight: 16238 Da
        10         20         30         40         50         60
         |          |          |          |          |          |
 MRSLLILVLC FLPLAALGKV FGRCELAAAM KRHGLDNYRG YSLGNWVCAA KFESNFNTQA

        70         80         90        100        110        120
         |          |          |          |          |          |
 TNRNTDGSTD YGILQINSRW WCNDGRTPGS RNLCNIPCSA LLSSDITASV NCAKKIVSDG

       130        140     147
         |          |       |
 NGMNAWVAWR NRCKGTDVQA WIRGCRL


ORIGIN
        1 mrsllilvlc flplaalgkv fgrcelaaam krhgldnyrg yslgnwvcva kfesnfntqa
       61 tnrntdgstd ygilqinsrw wcndgrtpgs rnlcnipcsa llssditasv ncakkivsdg
      121 ngmsawvawr nrckgtdvqa wirgcrl
//
```

Figure 14.8 Database file formats for amino acid sequence
PIR (upper), Swiss-Prot (middle) and GenPept (lower) formats for amino acid sequence of chicken egg-white lysozyme are represented. Swiss-Prot format is the default format of EMBL/EBI and GenPept format is the default format adopted by GenBank (Entrez) and DDBJ (DBGet). Fasta format is shown in subsection 5.1.2.

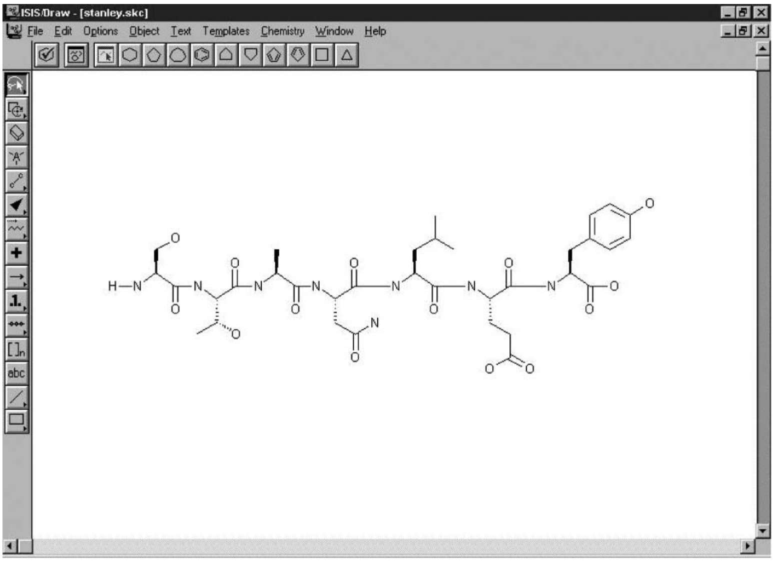

ISIS/Draw - [stanley.skc]
File Edit Options Object Text Templates Chemistry Window Help
H–N
abc

Figure 14.9 Two-dimensional structure sketch with ISIS Draw
The 2D structure of a heptapeptide is sketched by importing the sequence file in text format, STANLEY.

The full atomic representations of the 3D structures constitute 3D graphics. The common atomic coordinate files for the 3D biomacromolecular structures is PDB (Protein Data Bank) format (Figure 14.10). which is recognizable by most molecular graphics/modeling software programs. The pdb files of biomaromolecules can be retrieved from Research Collaboratory for Structural Bioinformatics (RCSB) at http://www.rcsb.org/pdb/ (Bourne *et al*., 2004) and visualized (Figure 14.11) with freeware programs such as RasMol (Sayle and Milner-White, 1995) Cn3D (Wang *et al*., 2000) or KineMage (Richardson and Richardson, 1994), retrievable from http://www.umass.edu/microbio/rasmol/,http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.html or http://orca.st.usm.edu/~rbateman/kinemage/ respectively.

14.4.4 Resource: Internet

The Internet is a network of networks meaning that many different networks operated by a multitude of organizations are connected together collectively to the Internet (Chellen, 2000; Crumlish, 1999; Swindell *et al*., 1996; Wyatt, 1995). The term ‘Internet’ derives from connecting networks, technically known as internetworking. It is based on a so-called ‘packet-switching network’ whereby data are broken up into packets. These packets are forwarded individually by adjacent computers on the network, acting as routers, and are reassembled in their original form at their destination. Packet switching allows multiple users to send information across the network both efficiently and simultaneously. Because packets can take alternate routes through the network, data transmission is easily maintained if parts of the network are damaged or not functioning efficiently. The widespread use of Transmission Control Protocol (TCP) together with Internet Protocol (IP) allows many networks to become interconnected through devices call gateways. Each computer connected to the Internet has a unique IP number or IP address consisting of four sets of 8-bit numbers (0–255) separated by a period. It provides a unique identifier that allows computers to communicate globally over the Internet. If a user is unsure of the IP address

```
HEADER    HYDROLASE (O-GLYCOSYL)                  01-FEB-75   1LYZ      1LYZ
COMPND    LYSOZYME (E.C.3.2.1.17)                                       1LYZ
SOURCE    HEN (GALLUS $GALLUS) EGG WHITE                                1LYZ
... ...
... ...
ATOM    258   N   PHE    34       2.966  25.379  11.370  7.00  1.50      1LYZ
ATOM    259   CA  PHE    34       3.955  25.906  11.749  6.00  1.50      1LYZ
ATOM    260   C   PHE    34       4.614  25.247  12.887  6.00  1.50      1LYZ
ATOM    261   O   PHE    34       5.801  25.379  12.887  8.00  1.50      1LYZ
ATOM    262   CB  PHE    34       3.494  27.356  12.002  6.00  1.50      1LYZ
ATOM    263   CG  PHE    34       3.362  27.949  10.612  6.00  1.50      1LYZ
ATOM    264   CD1 PHE    34       2.241  29.004  10.297  6.00  1.50      1LYZ
ATOM    265   CD2 PHE    34       4.614  27.949   9.728  6.00  1.50      1LYZ
ATOM    266   CE1 PHE    34       1.978  29.663   9.160  6.00  1.50      1LYZ
ATOM    267   CE2 PHE    34       4.417  28.674   8.402  6.00  1.50      1LYZ
ATOM    268   CZ  PHE    34       3.164  29.400   8.212  6.00  1.50      1LYZ
ATOM    269   N   GLU    35       3.889  24.653  13.771  7.00  1.50      1LYZ
ATOM    270   CA  GLU    35       4.680  23.796  14.971  6.00  1.50      1LYZ
ATOM    271   C   GLU    35       5.273  22.412  14.529  6.00  1.50      1LYZ
ATOM    272   O   GLU    35       6.526  22.149  14.845  8.00  1.50      1LYZ
ATOM    273   CB  GLU    35       4.087  23.599  15.982  6.00  1.50      1LYZ
ATOM    274   CG  GLU    35       3.757  23.665  17.308  6.00  1.50      1LYZ
ATOM    275   CD  GLU    35       4.680  24.522  17.940  6.00  1.50      1LYZ
ATOM    276   OE1 GLU    35       5.603  24.588  17.687  8.00  1.50      1LYZ
ATOM    277   OE2 GLU    35       4.153  25.642  18.509  8.00  1.50      1LYZ
ATOM    278   N   SER    36       4.680  21.687  13.834  7.00  1.50      1LYZ
ATOM    279   CA  SER    36       4.878  20.435  13.455  6.00  1.50      1LYZ
ATOM    280   C   SER    36       5.142  20.039  11.876  6.00  1.50      1LYZ
ATOM    281   O   SER    36       5.669  18.919  11.370  8.00  1.50      1LYZ
ATOM    282   CB  SER    36       4.285  19.446  14.024  6.00  1.50      1LYZ
ATOM    283   OG  SER    36       2.966  19.116  13.266  8.00  1.50      1LYZ
... ...
TER    1002       LEU   129                                              1LYZ
HETATM 1003   O   HOH     1       1.437  16.676  19.902  7.36  1.59   1  1LYZ
HETATM 1004   O   HOH     2       -.616  11.133  19.523  8.12  1.80   2  1LYZ
... ...
CONECT   48  47  981                                                     1LYZ
... ...
END                                                                      1LYZ
```

Figure 4.10 PDB file (partial) for 3D structure of hen's egg-white lysozyme (1LYZ.pdb) The abbreviated file shows partial atomic coordinates for residues 34–36. Informational lines such as AUTHOR (contributing authors of the 3D structure), REVDAT, JRNL (primary bibliographic citation), REMARK (other references, corrections, refinements, resolution and missing residues in the structure), SEQRES (amino acid sequence), FTNOTE (list of possible hydrogen bonds), HELIX (initial and final residues of α-helices), SHEET (initial and final residues of β-sheets), TURN (initial and final residues of turns, types of turns), and SSBOND (disulfide linkages) are deleted here for brevity. Atomic coordinates for amino acid residues are listed sequentially on ATOM lines. The following HETATM lines list atomic coordinates of water and/or ligand molecules.

for his/her computer, it can be obtained automatically by logging on to a Web site (www.ip-address.com/). IP addresses are powerful identifiers for computers, but they are cumbersome for human users. The domain name (e.g. ncbi.nlm.nih.gov) provides an optional identifier on the Internet that is more intuitive than IP addresses and easier to remember, because the domain name often describes the nature of academic institution or business that holds the domain. The name of a computer can then be <computer>.domain with six categories of top-level domain names in the US. Outside the US., the top-level domain names are replaced with a two-letter code specifying the country (Table 14.9).

If your computer at work is on a local area network (LAN), you may already have access to an external connection. Almost all universities have permanent connections

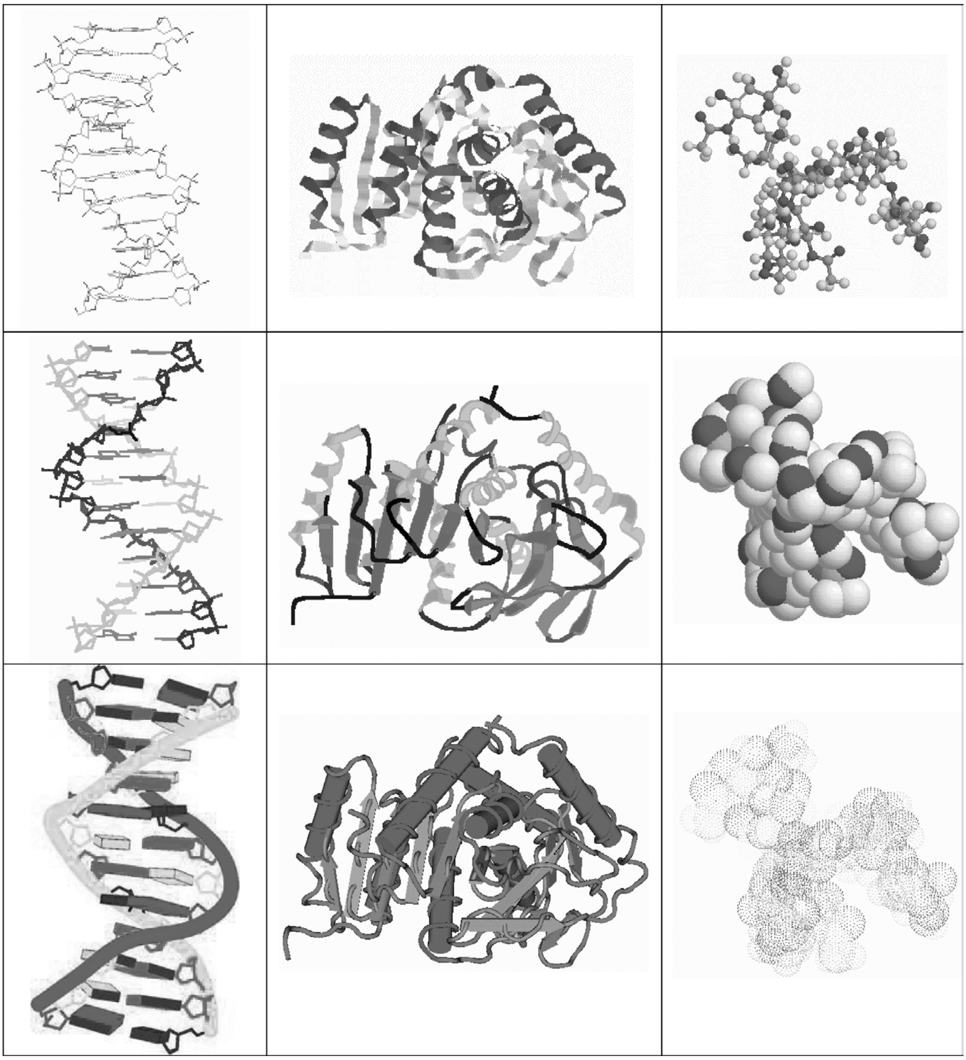


Figure 14.11 Various formats for visualizing 3D structures of biomolecules
The first column displays base pairing features of DNA (5′-dACCGACGTCGGT-3′, 424D.pdb) in hydrogen bonded wireframe, sticks with colored bases and cartoon representations. The second column illustrates the secondary structural features of protein (lactate dehydrogenase, 1LDB.pdb) in ribbon, ribbon-arrow, and cylinder-arrow formats. The third column depicts all-atom models of glycan (GN2bMa3(GN2bMa6)Mb4GNb4GN, Linucs1294.pdb) in ball-stick, spacefill and dot representations.

between their internal systems and the Internet. If you do not have access to an institutional connection, you have to obtain an access from an Internet service provider (ISP). For a full Internet connection, you must have software to make your computer use the Internet protocol, TCP/IP, and if your connection is through a modem this software will probably use point-to-point protocol (PPP).

14.4.4.1 List servers and newsgroups. A list server/mail server contains discussion group created to share ideas and knowledge on a subject; LISTSERV is the most common list server program. A message sent to a list is copied and then forwarded by

TABLE 14.9 Top-Level Domain Names

Inside the US	
.com	Commercial site
.edu	Educational site
.gov	Government site
.mil	Military site
.net	Gateway or network host
.org	Organization site
Outside the US, e.g.	
.ca	Canada
.ch	Switzerland
.cn	China
.de	Germany
.dk	Denmark
.es	Spain
.fr	France
.in	India
.it	Italy
.jp	Japan
.ru	Russia
.uk	Britain

Note: .com and .org have been used both inside and outside the US.

e-mail to every person who subscribes to the list, thereby providing an excellent resource for distributing information to a group with a shared interest. Discussion groups are usually created and sometimes monitored by someone with an interest in that subject. You join the list by sending an appropriately worded e-mail request to the list. The program automatically reads your e-mail message, extracts your address and adds you to the circulation list.

Unlike list servers, which disseminate information on a specific topic from one person to many, newsgroup servers (e.g. USENET) provide access to thousands of topic-based discussion group services that are open to everyone. Access is provided through a local host or news server machine.

14.4.4.2 Telnet. Telnet is a program using the communication protocol of the Internet (TCP/IP) to provide a connection on to remote computers. You can use Telnet to contact a host machine simply by typing in the host name or IP number if you have Internet access from your computer. You will then be asked for a login identity and your password. Often buried within Telnet is a version of FTP, so you can transfer files from the TCP/IP host to your own computer.

14.4.4.3 World Wide Web. The World Wide Web (WWW) is the world wide connection of computer servers and a way of using the vast interconnected network to find and view information from around the world (Bullock, 2003; Stout, 1996). Internet uses a language, TCP/IP for talking back and forth. The TCP part determines how to take apart a message into small packets that travel on the Internet and then reassemble them at the other end. The IP part determines how to get to other places on the Internet. The WWW uses an additional language called the HyperText Transfer Protocol (HTTP). The main use of the Web is for information retrieval, whereby multimedia documents are copied for local viewing. Web documents are most commonly written in HyperText Markup Language (HTML), which describes where hypertext links are located within the document. These hyperlinks provide connections between documents, so that a simple click on a hypertext word or picture on a Web page allows your computer to extend across the Internet and bring the document to your computer. The central repository having information the user wants, is called a server and your (user's) computer is a client of the server. Because HTML is predominantly a generic descriptive language based around text, it does not define any graphical descriptors. The common solution was to use bit-mapped graphical images in so-called GIF (graphical interchange format) or joint photographic experts group (JPEG) format.

Every Web document has a descriptor, Uniform Resource Locator (URL), which describes the address and file name of the page. The key to the Web is the browser program, which is used to retrieve and display Web documents. The browser is an Internet compatible program and does three things for Web documents:

1. It uses the Internet to retrieve documents from servers.
2. It displays these documents on your screen, using formatting specified in the document
3. It makes the displayed documents active.

The common browsers are Netscape Navigator from Netscape Communications (http://channels.netscape.com/ns/browsers/default.jsp) and Internet Explore from Microsoft (http://www.microsoft.com/windows.ie/default.mspx). Other browsers include Opera (http://www.opera.com) and Mozilla (http://www.mozilla.org), which is an open source browser (publicly shared code so that other programmers can contribute to improvement).

To request a Web page on the Internet, you either click your mouse on a hyperlink or type in the URL. The HTML file for the page is sent to your computer together with each graphic image, sound sequence or other special effect file that is mentioned in the HTML file. Since some of these files may require special programming that has to be added to your browser, you may have to download the program the first time you receive one of these special files. These programs are called helper applications, add-ons or plug-ins.

A mechanism called Multipurpose Internet Mail Extensions (MIME), which allows a variety of standard file formats to be exchanged over the Internet using electronic mail has been adopted for use with WWW. When a user makes a selection through a hyperlink within an HTML document, the client browser posts the request to the designated web server. Assuming that the server accepts the request, it locates the appropriate file(s) and sends it with a short header at the top of each datafile/document to the client, with the relevant MIME header attached. When the browser receives the information, it reads the MIME types such as text/html or image/gif. The browsers have been built in such a way that they can simply display the information in the browser window. For a given MIME type, a local preference file is inspected to determine what (if any) local program (known as a helper application or plug-in) can display the information, and this program is then launched with the data file and the results are displayed in a newly opened application window. The important aspect of this mechanism is that it achieves the delivery of semantic content to the user, who can specify the style in which it will be displayed via the choice of an appropriate application program. For molecular visualization on many web sites, Chime plug-in (http://www.mdlchime.com) must be installed.

Once connected to the Web, a variety of WWW directories and search engines are available for those using the Web in a directed fashion. First, there are Web catalogues;

the best known of these is Yahoo (http://www.yahoo.com), which organizes Web sites by subject classification. You can either scroll through these subject categories or use the Yahoo search engine. It also simultaneously forwards your search request to other search engines such as AltaVista and Excite. Alternatively, there are the Web databases, where the contents of Web pages are indexed and become searchable such as InfoSeek (http://www.infoseek.com) and Lycos (http://www.lycos.com). Google (http://www.google.com) is extremely comprehensive. Pages are ranked based on how many times they are linked from other pages, thus a Google search would bring you to the most well-traveled pages that match your search topic. HotBot (http://www.hotbot.com) is relatively comprehensive and regularly updated. It offers form-based query tools that eliminate the need for you to formulate query statements. Many other specialized search engines can be found in EasySearcher 2 (http://www.easy-searcher.com/ez2.html).

14.4.4.4 File fetching: File Transfer Protocol.

The File Transfer Protocol (FTP) allows you to download (get/receive) resources from a remote computer on to your own, or upload (put/send) these from your computer on to a remote computer. Sometimes FTP access to files may be restricted. To retrieve files from these computers, your must know the address, and have a user ID and a password. However, many computers are set up as anonymous FTP servers, where the user usually logs in as anonymous and give his/her e-mail address as a password. Internet browsers such as Netscape Navigator and Internet Explorer support anonymous FTP. Simply change the URL from http:// to ftp:// and follow it with the name of the FTP site you wish to go to. However, you must fill in your identity under the Options menu so that the browser can log you in as an anonymous user. The program and files on FTP sites are usually organized hierarchically in a series of directories. Those on anonymous FTP sites are often in a directory called pub (i.e., public). It is worth remembering that many FTP sites are running on computers with a UNIX operating system (Reichard and Foster-Johnson, 1999) that is case sensitive.

Some of the important commands in the FTP environment are:

Command	Action
help	Print out information on a specific command
ls	List the contents of the directory on the remote host
cd	Change the working directory on the remote host
lcd	Change the working directory on the local host
get/mget	Copy a single file (multiple files) from the remote host to the local host
put/mput	Copy a single file (multiple files) from the local host to the remote host
binary	Change the file-transfer mode to binary
ascii	Change the file-transfer mode to ASCII
prompt	Toggle the interactive mode that ask you to confirm every transfer

Many FTP servers supply text information when you login, in addition to the readme file. You can get help at the FTP prompt by typing ‘help’ or ‘?’. Before you download/upload the image file, select the option of transferring from *asc* (ascii) for text to *bin* (binary) for the image (graphic) files. Most files are stored in a compressed or zipped format. Some programs for compressing and uncompressing come as one integrated package, while others are two separate programs. The most commonly used compression programs are as follows:

Suffix	Compression program
.sit	Mac StuffIt
.sea	Mac self-extracting archive
.zip	Win PkZip, WinZip
.z	UNIX compress
.tar	UNIX tape

14.4.4.5 Internet vs. Intranet. The Web of the Internet is a way to communicate with people at a distance around the globe but the same infrastructure can be used to connect people within an organization known as the Intranet. Such Intranets provide an easily accessible repository of relevant information (but isolated from the world through firewalls), capitalizing on the simplicity of the Web interface. They also provide an effective channel for internal announcements or confidential communications within the organization.

The Internet has an open architecture network made possible by standards that apply and allow different proprietary technologies to work together. The organizations that pull together from a community of Internet users and work together for a perceived benefit to the Internet and WWW include Internet Society (ISOC) and World Wide Web Consortium (W3C). ISOC (http://www.isoc.org) is the international organization responsible for cooperation and coordination on the Internet. It acts to hold together a host of groups responsible for Internet infrastructure standards and protocols. W3C (http://www.w3c.org) functions as an archive of Web information for developers and users. The Consortium also promotes standards and develops prototype applications to demonstrate the use of new Web technologies. It also works in accordance with ISOC to help develop and maintain issues that can impact the evolution of the Web.

14.3.5 Internet resources of biochemical interest

The Internet has become the major information resources for biological sciences (Duscart, 2002). One of the most valuable web resources for scientific literatures is PubMed (www.ncbi.nlm.nih.gov/PubMed/), which is accessible from Entrez at http://www.ncbi.nlm.nih.gov/entrez. PubMed is provided by the National Library of Medicine, which is part of the National Institute of Health (NIH). Many well-regarded journals in biochemistry, cell biology, molecular biology and related fields as well as clinical publications of interest to medical professionals are indexed in PubMed. The resource can be searched by a keyword entry with the Boolean operators (AND, OR and NOT). Users can specify which database fields to check for each search term. The Preview/Index menu allows the user to build a detailed query interactively. The clipboard allows the user to collect and save individual search results.

The first issue for each volume of *Nucleic Acids Research* (http://nar.oupjournals.org/) since 1996 (vol. 24) has been reserved for presenting molecular biology databases. The Web server issue was added in 2003 (vol. 31). Separate Database and Web server issues are published as supplements 01 and 02 respectively since 2004 (Vol. 32). Relevant database resources for biochemistry/molecular biology have been recently compiled/updated (Galperin, 2004; Galperin, 2006). The resources available on the Internet are always changing. Some servers may move their URL addresses or become non-operational while new ones may be created online. Therefore it is not the goal of this text to provide a nearly complete guide to the WWW databases and servers, but instead to provide samples of bioinformatics related resources of general utilities (Table 14.10).

TABLE 14.10 Some Web sites of general biochemical interest

WWW Site	URL
Intl. Union of Biochem. Mole. Biol. (IUBMB)	http://www.chem.qmw.ac.uk/iubmb/
National Biotechnology Information Facility	http://www.nbif.org/data/data.html
EBI Biocatalog	http://www.ebi.ac.uk/biocat/
Dbcat	http://www.infobiogen.fr/services/dbcat/
Bioinformatic WWW sites	http://biochem.kaist.ac.kr/bioinformatics.html
Bioinformatics links directory	http://www.bioinformatics.ubc.ca/resoruces/links_directory
Bioinformatic tools	http://www.embl-heidelberg.de/Services/index.html
Bioinformatic search	http://www.hgsc.bcm.tmc.edu/Search/Launcher/
INFOBIOGEN Catalog of DB	http://www.infobiogen.fr/services/dbcat/
Links to other bio-Web servers	http://www.gdb.org/biolinks.html
Survey of Molecular Biology DB and servers	http://www.ai.sri.com/people/mimbd/
Human genome project information	http://www.ornl.gov/TechResources/Human_Genome
IUBio archive	http://iubio.bio.indiana.edu/soft/molbio/Listing.html
Structure Biology software DB	http://ks.uiuc.edu/Development/biosoftdb

For a comprehensive catalog of biochemical databases, Dbcat (Discala *et al*., 2000) at http://www.infobiogen.fr/services/dbcat/ should be consulted.

The National Center for Biotechnology Information (NCBI) (Wheeler *et al*., 2001) was established in 1988 as a division of the National Library of Medicine (NLM) of the National Institutes of Health (NIH) in Bethesda, Maryland (http://www.ncbi.nlm.nih.gov/). Its mandate is to develop new information technology to aid our understanding of the molecular and genetic processes that underlie health and disease. The specific aims of NCBI include:

a) the creation of automated systems for storing and analyzing biological information;

b) the development of advanced methods of computer-based information processing;

c) the facilitation of user access to databases and software; and

d) the co-ordination of efforts to gather biotechnology information worldwide.

Entrez (http://www.ncbi.nlm.nih.gov/entrez) is the comprehensive, integrated retrieval server of biochemical information and serves as the entry point to NCBI resources (Wheeler *et al*., 2006).

The European Molecular Biology network (EMBnet) was established in 1988 to link European laboratories that used biocomputing and bioinformatics in molecular biology research by providing information, services and training to users in European laboratories, through designated sites operating in their local languages. EMBnet consists of National sites (e.g. INFOBIOGEN, GenBee, SEQNET), Specialist site, the European Bioinformatics Institute (EBI) in the UK (Brooksbank *et al*., 2005) and Associate sites. National sites, which are appointed by the governments of their respective nations, have mandate to provide databases, software and online services. Special sites are academic, industrial or research centers that are considered to have particular knowledge of specific areas of bioinformatics responsible for the maintenance of biological database and software. Associate sites are biocomputing centers from non-European countries. The Sequence Retrieval System (SRS) (Etzold *et al*., 1996) developed at European Bioinformatics Institute (EBI) at http://www.ebi.ac.uk/ is a network browser for databases in molecular biology allowing users to retrieve, link and access entries from all the interconnected resources. The system links nucleic acid, EST, protein sequence, protein pattern, protein structure, specialist and/or bibliographic databases.

The DNA Data Bank of Japan (http://www.genome.ad.jp/) is a Japanese network of database and computational service for genome research and related areas in molecular and cellular biology (Tateno *et al*., 2005). It is operated jointly by the Institute for Chemical Research, Kyoto University and the Human Genome Center of the University of Tokyo.

German Cancer Research Center (DKFZ) in Heidelberg maintains Glycosciences (http://www.glycosciences.de/index.php), which provides glycomic databases, modeling and tools to serve the need of glycoscientists. The glycan information is also available at The Consortium for Functional Glycomics (http://www.functionalglycomics.org/).

14.5 INFORMATICS

The amount of information in biological sequences is related to their compressibility. Conventional text compression schemes are so constructed that the original data can be recovered perfectly without losing a single bit. Text compression algorithms are designed to provide a shorter description in the form of a less redundant representation, normally called a code, which may be interpreted and converted back into the uncompressed message in a reversible manner (Rival *et al*., 1996). Biochemistry is full of such code words (e.g. A, C, G, T/U for nucleotides, A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y for amino acids and A, B, E, F, G, M, N, R, X for glycoses) to provide genome/proteome/glycome information. Thus the nucleic acid/protein/glycan sequences are reduced neatly to character strings in which a single letter (code word) represents a single nucleotide/amino acid/glycose. The challenge in representing/managing sequence data is verification of the correctness of the data via annotation of data, their effective search/retrieval via efficient search engine and accurate/intuitive presentation of retrieved data.

14.5.1 Introduction to database

The data are the unevaluated, independent entries that can be either numeric or non-numeric e.g., alphabetic or symbolic. Information is ordered data that are retrieved according to a user's need. If raw data are to be effectively processed to yield information, they must first be organized logically into files. In computer files, a field is the smallest unit of data, which contains a single fact. A set of related fields is grouped as a record, which contains all the facts about an item in the table, and a collection of records of the same type is called a file, which contains all facts about a topic in the table. Databases are electronic filing cabinets that serve as a convenient and efficient means of storing vast amounts of information.

A database system is a computerized information system for the management of data by means of a general purpose software package called a database management system (DBMS). Operationally, a database (databank) is a collection of interrelated files created with a DBMS (Mullins, 2002). The content of a database is obtained by combining data from all the different sources in an organization so that data are available to all users such that redundant data can be eliminated or at least minimized. Figure 14.12 illustrates the relationship among the components of a typical database. Two files are interrelated (logically linked) if they contain common data types that can be interpreted as being equivalent. Two files are also interrelated (physically linked) if the value of a data type of one

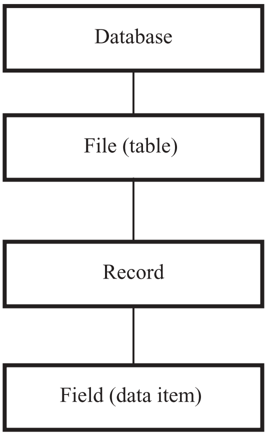


Figure 14.12 Relationship among components of databases

file contains the physical address of a record in the other file. Relational database stores information in a collection of files, each containing data about one subject. Because the files are logically linked, we can use information from more than one file at a time.

Formally, a database consists of three components, i.e. data in a searchable form, generic search engine and query language appropriate to the domain of data. The entire database is composed of entries that are discrete coherent parcels of information indexed by a set of pointers for efficient search. The search engine is responsible for:

- the translation of the user's query into the encoded space of the data;
- preliminary screening of the query against database keys;
- pattern matching of the query; and
- presentation of the resultant hits.

14.5.2 Biochemical databases

Many biological databases (databanks) are embedded with tutorials that make it easy to explore their facilities. There are three sources of biological databases; in-house dedicated sources (private and limited for focused projects), databases assembled by companies (mainly fees for services: extensive and high-quality but expensive and restrictive such as Celera Genomics and Incyte Genomics), and public databases (such as GenBank, EMBL and DDBJ). An important distinction exists between primary (archival) and secondary (curated) databases. The primary databases represent experimental results with some interpretation (Table 14.11). Their record is the sequence or structure as it was experimentally derived.

Primary data collections related to biomolecules include:

- Nucleic acid sequences
- Amino acid sequences of proteins
- Glycan sequences/structures of glycoproteins/glycolipids
- Protein and nucleic acid structures
- Expression patterns/functions of genes
- Protein functions
- Publications

TABLE 14.11 Some primary database sites of biochemical interest

Database	URL
Comprehensive:	
NCBI: Sequence & structure resources	http://www.ncbi.nlm.nih.gov
EBI: Data and tool resources	http://www.ebi.ac.uk
Sequence, nucleic acids:	
GenBank	http://www.ncbi.nlm.nih.gov/Web/Genbank/
European Bioinformatics Institute (EBI)	http://www.ebi.ac.uk/
DNA Database of Japan (DDBJ)	http://www.ddbj.nig.ac.jp/
Sequence, proteins:	
Protein Information Resource (PIR), Internat'l	http://pir.georgetown.edu/
Swiss-Prot	http://expasy/hcuge.ch/sprot/
Munich Inform. Center for Protein Sequ. (MIPS)	http://www.mips.biochem.mpg.de/
UniProt	http://www.uniprot.org/
Structure, biomacromolecules:	
Protein Data Bank at RCSB	http://www.rcsb.org/pbd/
Nucleic acid database (NDB)	http://ndbserver.rutgers.edu/NDB/
EBI-MSD: macromolecular structure database	http://msd.ebi.ac.uk/
Protein quaternary structure	http://pqs.ebi.ac.uk/
Glycosciences, SweetDB	http://www.glycosciences.de/sweetdb/index.php
Consortium for Functional Glycomics, GlycanDB	http://www.functionalglycomics.org/molecule/jsp/carbohydrate/carboMoleculeHome.jsp
Structure, organic/biomolecules:	
ChEBI: Biomolecules	http://www.ebi.ac.uk/chebi/
Cambridge Structural Database (CSD)	http://www.ccdc.cam.ac.uk/prods/csd/csd.html
Klotho: classification of biochem. compounds	http://www.biocheminfo.org/klotho
PubChem: bioactivities of org. compounds	http://pubchem.ncbi.nlm.nih.gov/
Super Natural: Natural compounds	http://bioinformatics.charite.de/supernatural
Expression:	
2DPAGE	http://www.expasy.ch/ch2d/ch2d-top.html
Microarray	http://industry.ebi.ac.uk/arrayexpress
Biochemical pathways	http://www.genome.ad.jp/kegg/

Sequence databases generally specialize in one type of sequence data, i.e. DNA, RNA or protein (Higgins and Taylor, 2000). Structure data must unambiguously define the atomic connectivities and the precise three-dimensional coordinates of all atoms within the molecule. These sequences and structures are the items to be computed on and worked with as the valuable components of the primary databases. Generally, the gateways to sequence and structure databases include:

- Retrieval of sequences/structures from the database;
- Sequence search, comparison/alignment;
- Translation of DNA sequences to protein sequences or *vice versa* in reverse translation;
- Simple structure analysis and prediction;
- Pattern recognition;
- Multiple alignment and phylogenetic relationships;
- Molecular graphics.

The secondary databases contain the fruits of analyses of the sequences or structures in the primary sources such as patterns, motifs, functional sites and so on. Many databases known as boutique (specialized) databases select, annotate and recombine data focused on particular topics, and include links affording streamlined access to information about subjects of interest.

14.5.3 Database retrieval

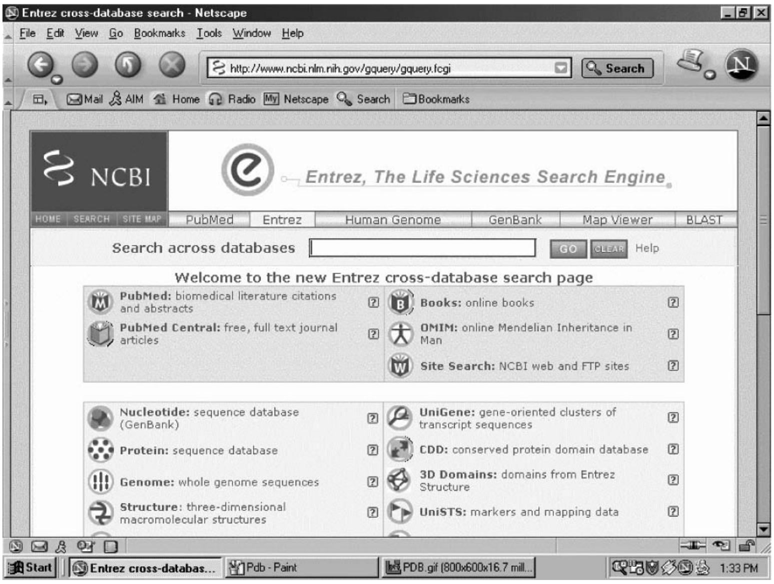
Most biochemical and/or molecular biology databases in the public domains are flat-file databases. Each entry of a database is given a unique identifier, i.e. an entry name and/or accession number so that it can be retrieved uniformly by the combination of the database name and the identifier.

Two comprehensive integrated retrieval systems for nucleic acids and proteins, Entrez of NCBI, and SRS of EBI will be described briefly. The molecular biology database and retrieval system, Entrez (Schuler *et al*., 1996) was developed at and maintained by NCBI of NIH to allow retrieval of biochemical data and bibliographic citation from its integrated databases. Entrez typically provides access to:

- DNA sequences (from GenBank, EMBL and DDBJ);
- protein sequences (from PIR, SWISS-PROT, PDB);
- genome and chromosome mapping data;
- 3D structures (from PDB); and
- PubMed bibliographic database (MEDLINE).

Entrez searches can be performed using one of two Internet-based interfaces. The first is a client-server implementation known as NetEntrez. This makes a direct connection to an NCBI computer. Because the client software resides on the user's machine, it is up to the user to obtain, install and maintain the software, downloading periodic updates as new features are introduced. The second implementation is over the World Wide Web and is known as WWW Entrez or WebEntrez (simply referred to as Entrez). This option makes use of available Web browsers (e.g., Netscape or Explorer) to deliver search results to the desktop. The Web allows the user to navigate by clicking on selected words in an entry. Furthermore, the Web implementation allows for the ability to link to external data sources. While the Web version is formatted as sequential pages, the Network version uses a series of windows with faster speed. The NCBI databases are, by far the most often accessed by biochemists and some of their searchable fields include plain text, author name, journal title, accession number, identity name (e.g., gene name, protein name, chemical substance name), EC number, sequence database keyword and medical subject heading.

The Entrez home page (Figure 14.13) is opened by keying in URL, http://www.ncbi.nlm.nih.gov/Entrez/. An Entrez session is initiated by choosing one of the available databases (e.g., click Nucleotides for DNA sequence, Proteins for protein sequence or 3D structures for the 3D coordinates) then composing a Boolean query designed to select a small set of documents after choosing Primary accession for the Search Field. The term comprising the query may be drawn from either plain text or any of several more specialized searchable fields. Once a satisfactory query is selected, a list of summaries (brief descriptions of the document contents) may be viewed. Successive rounds of neighboring (to related documents of the same database types) or linking (to associate records in other databases) may be performed using this list. After choosing an appropriate report format (e.g., GenBank, fasta), the record may be printed and saved.


Entrez cross-database search - Netscape
File Edit View Go Bookmarks Tools Window Help
http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi
Search
Mail AIM Home Radio Netscape Search Bookmarks
NCBI
Entrez, The Life Sciences Search Engine
HOME SEARCH SITE MAP
PubMed Entrez Human Genome GenBank Map Viewer BLAST
GO CLEAR Help
Start Entrez cross-databas... Pdb - Paint PDB.gif (800x600x16.7 mill... 1:33 PM

Search across databases

Welcome to the new Entrez cross-database search page

PubMed: biomedical literature citations and abstracts

PubMed Central: free, full text journal articles

Books: online books

OMIM: online Mendelian Inheritance in Man

Site Search: NCBI web and FTP sites

Nucleotide: sequence database (GenBank)

Protein: sequence database

Genome: whole genome sequences

Structure: three-dimensional macromolecular structures

UniGene: gene-oriented clusters of transcript sequences

CDD: conserved protein domain database

3D Domains: domains from Entrez Structure

UniSTS: markers and mapping data

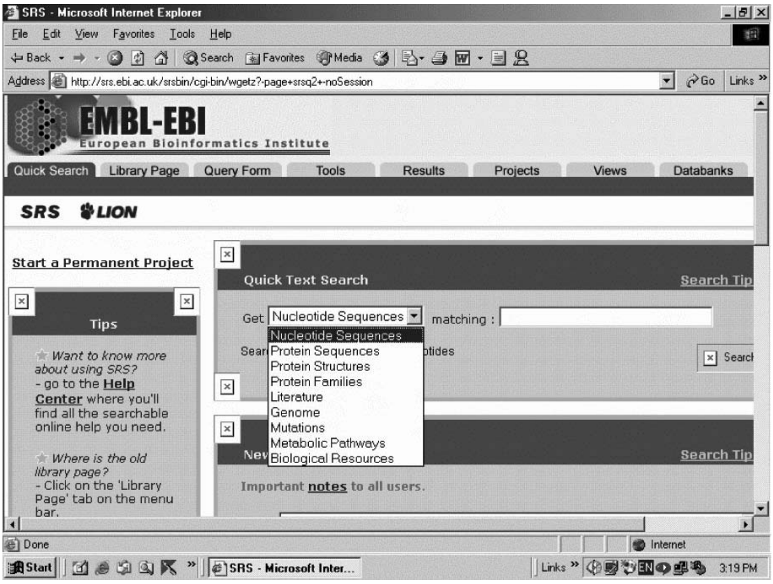
Figure 14.13 Comprehensive bioinformatic retrieval system of Entrez

One of the central activities of the European Bioinformatics Institute (EBI) (Emmert *et al*., 1994) is development and distribution of the EMBL nucleotide sequence database (Stoesser *et al*., 2001). This is a collaborative project with GenBank (NCBI, USA) and DDBJ (DNA Database of Japan) to ensure that all the new and updated database entries are shared between the groups on a daily bases. The search of sequence databases and an access to various application tools can be approached from the home page of EBI at http://www.ebi.ac.uk/. The Sequence Retrieval System (Etzold *et al*., 1996) is a network browser for databases at EBI and several databases. The system allows users to retrieve, link and access entries from all the interconnected resources such as nucleic acid, EST, protein sequence, protein pattern, protein structure, specialist/boutique and/or bibliographic databases. The SRS is also a database browser of DDBJ, ExPASy and a number of servers as the query system. The SRS can be accessed from EBI Tools server at http://www2.ebi.ac.uk/Tools/index.html or directly at http://srs.ebi.ac.uk/. The SRS (Figure 14.14) permits users to formulate queries across a range of different database types via a single interface in three different modes:

Quick search: This is the fastest way to generate a query. Click the checkbox to select databank(s), e.g. EMBL for nucleotide sequences or SWISS_PROT for amino acid sequences. Enter the query word(s), accession ID, or regular expression, then click the Go button to start the search.

Standard query: Select the databanks and click Standard button under Query forms. This opens the standard query form where the user is given choices of data fields to search, operator to use, wild card to append, entry type and result views. There are four textboxes with corresponding data field selectors. After entering query strings to textboxes and choosing data fields, select sequence formats (embl, fasta or genbank) then click the Submit Query button to begin the search.

Extended query: Select the databanks and click Extended button to open the extended query form. Enter query string to the Description field, name of the organism to the Organism field and data string (yyyymmdd) in the Date field. Click Submit Query button to initiate the search.


SRS - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+srsq2+-noSession
Go Links
EMBL-EBI
European Bioinformatics Institute
Quick Search
Library Page
Query Form
Tools
Results
Projects
Views
Databanks
SRS LION
Start a Permanent Project
Tips
Quick Text Search
Search Tip
Get
Nucleotide Sequences
matching :
Nucleotide Sequences
Protein Sequences
Protein Structures
Protein Families
Literature
Genome
Mutations
Metabolic Pathways
Biological Resources
Searc
Search Tip
Done
Internet
Start
SRS - Microsoft Inter...
Links
3:19 PM

Want to know more about using SRS? - go to the **Help Center** where you'll find all the searchable online help you need.

Where is the old library page? - Click on the 'Library Page' tab on the menu bar.

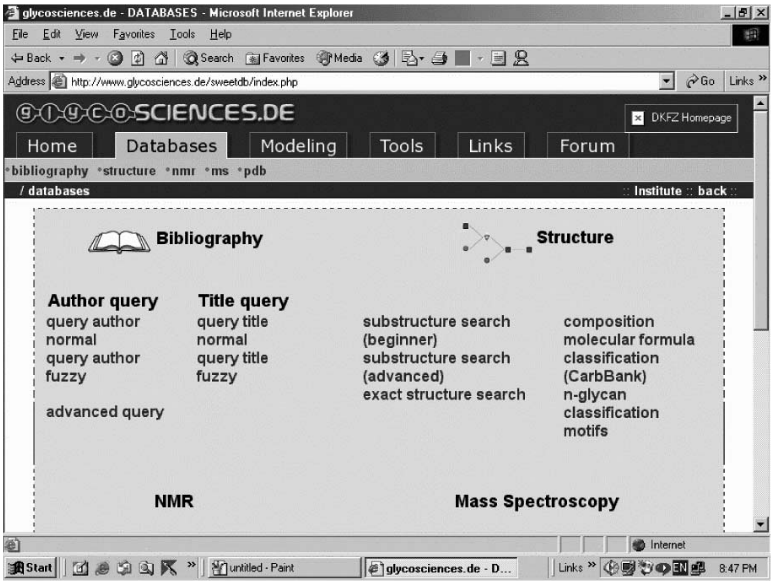
Important **notes** to all users.

Figure 14.14 Comprehensive retrieval system of SRS at EBI

The corresponding retrieval systems for glycan are Glycan Database of Consortium for Functional Glycomics (http://www.functionalglycomics.org) and Sweet Database of Glycosciences (http://www.glycosciences.de/index.php). The Consortium for Functional Glycomics (CFG), which is supported by the National Institute of General Medical Sciences offers access to glycan binding proteins (GBP) Database, glycosylation pathways (GT) Database and Glycan Database. The Glycan Database (http://www.functionalglycomics.org/molecule/jsp/carbohydrate/carboMoleculeHome.jsp) provides mechanism for search and retrieval of desired glycans. The Glycosciences, which is maintained by German Cancer Research Center in Heidelberg, offers bioinformatic tools (Modeling and Tools) and databases (Sweet Database) for glycobiology. Glycan structures can be searched and retrieved from the Database (http://www.glycosciences.de/sweetdb/index.php) using any of search options, e.g. substructure, composition, molecular formula, glycan (*N*-glycan) classification or motifs (Figure 14.15). After selecting the desired glycan from the returned list of hits, clicking Explore, LiGraph or NMR button leads respectively to theoretical 3D coordinate, molecular formula with relevant information and NMR spectrum of the glycan.

14.6 GENE ONTOLOGY

The genome era in biology has produced vast amounts of biological data accompanied by the widespread proliferation of biology-orientated databases. To make the best use of biological databases and the knowledge they contain, different kinds of information from different sources must be integrated that are understandable to biologists. An aspect of the integration effort is the development and use of annotation standards such as ontologies. Ontology is a domain of knowledge, represented by facts and their logical connections that can be understood by a computer for multiple purposes. Ontologies are becoming important in bioinformatics as they can articulate and makes generally accessible in a structured way the large amounts of biological knowledge, by linking to databases and being used for searching their contents. Ontologies are different from annotations in that


glycosciences.de - DATABASES - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.glycosciences.de/sweetdb/index.php
Go Links
GLYCOSCIENCES.DE
DKFZ Homepage
Home
Databases
Modeling
Tools
Links
Forum
•bibliography •structure •nmr •ms •pdb
/ databases
:: Institute :: back ::
Internet
Start
untitled - Paint
glycosciences.de - D...
Links
8:47 PM

Bibliography

Structure

Author query
query author normal
query author fuzzy

advanced query

Title query
query title normal
query title fuzzy

substructure search (beginner)
substructure search (advanced)
exact structure search

composition
molecular formula
classification (CarbBank)
n-glycan classification
motifs

NMR

Mass Spectroscopy

Figure 14.15 Glycan retrieval site of Sweet Database at Glycosciences

TABLE 14.12 Web reference sites for Gene Ontology

Web site	URL
AmiGo	http://www.godatabase.org/cgi-bin/go.cgi
Description logic	http://www.ida.liu.se/labs/ilslab/people/patla/DL/
GO project	http://www.geneontology.org/
GOA	http://www.ebi.ac.uk/GOA
GOBO	http://www.geneontology.org/doc/gobo.html
Graph	http://www.nist.gov/dads/HTML/graph.html
Gruber	http://www-ksl.stanford.edu/kst/what-is-an-ontology.html
GXD	http://www.informatics.jax.org/searches/expression_form.shtml
OMIM	http://www.ncbi.nlm.nih.gov/Entrez
Ontology definition	http://www.kr.org/top/
Ontology projects	http://www.cs.utexas.edu/users/mfkb/related.html

annotations are normally datafiles associated with items in a database whilst ontologies can be directly linked to the data in database with faster search capabilities.

Gene ontology (GO) is used to integrate genetic data about gene products with our knowledge of their properties. The GO catalogues this knowledge in three GO hierarchies; molecular function (the functions that they fulfill), biological process (the processes to which they contribute) and cellular component (their location within cells). The GO Consortium (http://www.geneontology.org/) is a collaborative effort to address information integration by providing consistent descriptors for gene products in different databases and standardizing classifications for sequences and sequence features (GO Consortium, 2006). In support of standardized GO nomenclature, the Gene Ontology Annotation (GOA) at (http://www.ebi.ac.uk/GOA) has organized, shared and integrate protein knowledgebase of UniProt (http://www.uniprot.org/) using the GO structured vocabulary (Camon *et al.*, 2004). Table 14.12 lists several GO and related Web sites.

14.7 REFERENCES

ALTSCHUL, S.F., GISH, W., MILLER, W. *et al.* (1990) *Journal of Molecular Biology*, **215**, 403–10.

ALTSCHUL, S.F., MADDEN, T.L., SCHAFFER, A.A. *et al.* (1997) *Nucleic Acids Research*, **25**, 3389–402.

ATTWOOD, T.K. and PARRY-SMITH, D.J. (1999) *Introduction to Bioinformatics*, Addision Wesley Longman Ltd. Harlow, UK.

BAXEVANIS, A.D. and OUELLETTE, B. (eds) (2005) *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, 3rd edn, John Wiley & Sons, Hoboken, NJ.

BANIN, E., NEUBERGER, Y., ALTSHULER, Y. *et al.* (2002) *Trends in Glycoscience Glycotechology*, **14**, 127–37.

BECKER, O.M., MACKERELL, A.D. JR., ROUX, B. and WATANABE, M. (eds) (2001) *Computational Biochemistry and Biophysics*, Marcel Dekker, Inc., New York.

BOURNE, P.E., ADDESS, K.J., BLUHM, W.F. *et al.* (2004) *Nucleic Acids Research*, **32**, D223–5.

BROOKSBANK, C., CAMERON, G. and THORNTON, J. (2005) *Nucleic Acids Research*, **33**, D46–53.

BROWN, T.A. (1994) DNA *Sequencing: A Practical Approach*, IRL Press, Oxford, UK.

BULLOCK, L. (2003) *The World Wide Web*, Raintree Steck-Vaughan, Austin, TX.

CAMON, E., MAGRANE, M., BARRELL, D. *et al.* (2004) *Nucleic Acids Research*, **32**, D262–6.

CARUTHERS, M.H. (1985) ‘Gene synthesis machine: DNA chemistry and its uses.’ *Science*, **230**, 281–5.

CHELLEN, S.S. (2000) *The Essential Guide to the Internet*, Routledge, New York.

CRUMLISH, C. (1999) *The Internet*, Sybex, San Francisco, CA.

DAYHOFF, M.O., SCHWARTZ, R.M. and ORCUTT, B.C. (1978), in *Atlas of Protein Sequence and Structure*, Vol. 5, Suppl. 3 (ed. M.O. Dayhoff), NBRF, Washington, DC, p 345.

DISCALA, C., BENIGNI, X., BARILLOT, E. and VAYSSEIX, G. (2000) *Nucleic Acids Research*, **28**, 8–9.

DURBIN, R., EDDY, S., KROGH, A. and MITCHISON, G. (1998) *Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids*, Cambridge University Press, Cambridge, UK.

DUSCART, G. (2002) *Biosciences on the Internet: A Student's Guide*, John Wiley & Sons, New York.

EISEN, M.B., SPELLMAN, T.P., BROWN, P.O. and BOTSTEIN, D. (1998) *Proceedings of the National Academy of Sciences, USA.*, **95**, 14863–8.

EMMERT, D.B., STOEHR, P.J., STOESSER, G. and CAMERON, G.N. (1994) *Nucleic Acids Research*, **22**, 3445–9.

ETZOLD, T., ULYANOV, A. and ARGO, P. (1996) *Methods in Enzymology*, **266**, 114–28.

FINDLEY, J.B.C. and GEISOW, M.J. (eds) (1989) *Protein Sequencing: A Practical Approach*, IRL Press/Oxford University Press, Oxford, UK.

GALPERIN, M.Y. (2004) *Nucleic Acids Research*, **32**, D3–D22.

GALPERIN, M.Y. (2006) *Nucleic Acids Research*, **34**, D3–D5.

GAUSCHIN, D., YERSHOV, G., ZASLAVSKY, A. *et al.* (1997) *Analytical Biochemistry*, **250**, 203–11.

GENE ONTOLOGY CONSORTIUM (2006) *Nucleic Acids Research*, **34**, D322–326.

HAUPT, K. and MOSBACH, K. (1998) *Trends in Biotechnology*, **16**, 468–75.

HENKEL, J.G. and CLARKE, F.H. (1985) *Molecular Graphics on the IBM PC Microcomputer*, Academic Press, Orlando, FL.

HIGGINS, D. and TAYLOR, W. (eds) (2000) *Sequence, Structure and Databanks: A Practical Approach*, Oxford University Press, Oxford, UK.

INNIS, M.A., GELFAND, D.H. and SNINSKY, J.J. (eds) (1999) *PCR Applications: Protocols for Functional Genomics*, Academic Press, San Diego.

JAMISON, D.C. (2003) *Perl Programming for Biologists*, Wiley-Liss, Hoboken, NJ.

JEANMOUGIN, F., THOMPSON, J.D., GOUY, M. *et al.* (1995) *Trends in Biochemical Science*, **23**, 403–5.

JOOS, T.O., STOLL, D. and TEMPLIN, M.F. (2002) *Current Opinions in Chemistry and Biology*, **6**, 76–80.

KAUFMAN, L. and ROUSSEEUW, P.J. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley, New York.

KIM, S.K., LUND J., KIRALY, M. *et al.* (2001) *Science*, **293**, 2087–92.

LESK, A.M. (2005) *Introduction to Bioinformatics*, 2nd edn, Oxford University Press, New York.

LIEBLER, D.C. (2002) *Introduction to Proteomics*, Humana Press, Totowa, N.J.

LIPMAN, D.J. and PEARSON, W.R. (1985) *Science*, **227**, 1435–41.

MACBEATH, G. and SCHRIEBER, S.L. (2000) *Science*, **289**, 1760–3.

MEYERS, M. and JERNIGAN, S. (2004) *Managing and Troubleshooting PCs*, McGraw Hill Technology Education, New York.

MCGINNIS, S. and MADDEN, T.L. (2004) *Nucleic Acids Research*, **32**, W20–5.

MUELLER, S. (2005) *Upgrading and Repairing PCs*. 16th edn, Que Corp. Indianapolis, IN.

MÜLLER, U.R. and NICOLAU, D.V. (eds) (2005) *Microarray Technology and its Applications*, Springer, Berlin, Germany.

MULLINS, C.S. (2002) *Database Administration: The Complete Guide to Practices and Procedures*, Addison-Wesley, Boston, MA.

NEEDLEMAN, S.B. and WUNSCH, C.D. (1970) *Journal of Molecular Biology*, **48**, 443–53.

PARK, S. and SHIN, I. (2002) *Angew. Chem. Int. Ed. Engl.*, **41**, 3180–2.

RAMSDEN, J. (2004) *Bioinformatics: An Introduction*, Kluwer Academic Publishers, Boston, MA.

REICHARD, K. and FOSTER-JOHNSON, E. (1999) *Teach Yourself UNIX*, 4th edn, IDG Books Worldwide, Foster City, CA.

Richardson, D.C. and Richardson, J.S. (1994) *Trends in Biochemical Science*, **19**, 135–8.

Rival, E., Dauchat, M., Delehaya, J.P. and Delgrange, O. (1996) *Biochimie*, **78**, 315–24.

Russell, R.B., Saqi, M.A., Sayle, R.A. *et al.* (1997) *Journal of Molecular Biology*, **269**, 423–39.

Sayle, R.A. and Milner-White, E.J. (1995) *Trends in Biochemical Science*, **20**, 374–6.

Schena, M. (ed.) (2000) *DNA Microarrays: A Practical Approach*, 2nd edn, Oxford University Press, Oxford, UK.

Schena, M. (2003) *Microarray Analysis*, John Wiley & Sons, Inc., New York.

Schuler, G.D., Epstein, J.A., Okawa, H. and Kans, J.A. (1996) *Methods in Enzymology*, **266**, 141–62.

Singer, M. and Berg, P. (1991) *Genes and Genomes: A Changing Perspective*, University Science Books, Mill Valley, CA.

Smith, T.F. and Waterman, M.S. (1981) *Journal of Molecular Biology*, **147**, 195–7.

Stekel, D. (2003) *Microarray Bioinformatics*, Cambridge University Press, Cambridge, UK.

Stoesser, G., Baker, W., van den Brock, A. *et al.* (2001) *Nucleic Acids Research*, **29**, 17–21.

Stout, R. (1996) *The World Wide Web Complete Reference*, Osborne McGraw-Hill, Berkeley, CA.

Swindell, S., Miller, R.R. and Myer, G. (eds.) (1996) *Internet for the Molecular Biologist*, Horizon Scientific Press, Washington, DC.

Takeuchi, T., Fukuma, D. and Matsui, J. (1999) *Analytical Chemistry*, **71**, 285–90.

Tateno, Y., Saitou, N., Okubo, K. *et al.* (2005) *Nucleic Acids Research*, **33**, D25–8.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) *Nucleic Acids Research*, **22**, 4673–80.

Tsai, C.S. (2002) *An Introduction to Computational Biochemistry*, John Wiley & Sons, New York.

Vaughan, T.J., Williams, A.W., Prichard, K. *et al.* (1996) *Nature Biotechnology*, **14**, 309–14.

Wang, Y., Geer, L.Y., Chappey, C. *et al.* (2000) *Trends in Biochemical Science*, **25**, 300–2.

Weber, S. (2004) *The Personal Computer*, Chelsea House Publishers, Philadelphia, PA.

Weininger, D. (1988) *Journal of Chemistry Information in Computer Science*, **28**, 31–6.

Wheeler, D.L., Church, D.M.C., Lash, A.E. *et al.* (2001) *Nucleic Acids Research*, **29**, 11–16.

Wheeler, D.L., Barrett, T., Benson, D.A. *et al.* (2006) *Nucleic Acids Research*, **34**, D173–D186.

Wyatt, A.L. (1995) *Success with Internet*, Boyd & Fraser, Danvers, MA.

World Wide Webs cited

BLAST server: http://www.ncbi.nlm.nih.gov/BLAST/
CarbBank: http://www.boc.chem.uu.nl/sugarbase/carbbank.html
Chime: http://www.mdlchime.com
ClustalW: ftp://ftp-igbmc.u-strasbg.fr/pub/
Cn3D: http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.html
Consortium for Functional Glycomics: http://www.functionalglycomics.org
DBcat: http://www.infobiogen.fr/services/dbcat
DDBJ: http://www.ddbj.nig.ac.jp
Dotplot: http://www.cgr.ki.se/cgr/groups/sonnhammer/Dotter.html
DOTTER: ftp://ftp.sanger.ac.uk/pub/dotter/
EasySearcher 2: http://www.easysearcher.com/ez2.html
EBI: http://www.ebi.ac.uk
EBI Tools: http://www2.ebi.ac.uk/Tools/index.html
EmBL: http:www.embl-heidelberg.de
Entrez (NCBI): http://www.ncbi.nlm.nih.gov/Entrez/
Expert Protein Analysis System (ExPASy): http://www.expasy.ch/
Free Software Foundation: http://www.fsf.org
GenBank: http://www.ncbi.nlm.nih.gov/Web/Genbank/
GenomeNet: http://www.genome.ad.jp/
GlycanDB: http://www.functionalglycomics.org/
GlycoWord: http://www.gak.co.jp/FCCA/glycoword/wordE.html
Glycosciences: http://www.glycosciences.de/index.php
GOA: http://www.ebi.ac.uk/GOA
GO Consortium: http://www.geneontology.org/
Google: http://www.google.com
HotBot: http://www.hotbot.com
InfoSeek: http://www.infoseek.com
IP address information: http://www.ip-address.com
ISIS Draw: http://www.mdli.com/download/isisdraw.html
Internet Society: http://www.isoc.org

KineMage	http://orca.st.usm.edu/~rbateman/kinemage/
Lycos:	http://www.lycos.com
Mozilla:	http://www.mozilla.org
Microarray e-library:	http://arrayit.com/e-library
Microsoft Explore:	http://www.microsoft.com/
NCBI:	http://www.ncbi.nlm.nih.gov
Netscape Communication:	http://home.netscape.com/
Nucleic Acids Research:	http://nar.oupjournals.org/
Open-Source Initiative:	http://www.opensource.org
Opera:	http://www.opera.com
Operon Database:	http://odb.Kuicr.kyoto-u.ac.jp/
Perl:	http://www.bioperl.org
PIR International:	http://pir.georgetown.edu
Protein Data Bank at RCSB:	http://www.rcsb.org/pbd/
PubMed:	www.ncbi.nlm.nih.gov/PubMed/
RasMol:	http://www.umass.edu/microbio/rasmol/
ReadSeq:	http://dot.imgen.bcm.tmc.edu:9331/seq-util/Options/readseq.html
SMILES:	http://www.daylight.com/dayhtml/smiles/
SRS:	http://srs.ebi.ac.uk/
SweetDB:	http://www.glycosciences.de/sweetdb/index.php
Swiss-Prot:	http://expasy/hcuge.ch/sprot/
UniProt:	http://www.uniprot.org/
World Wide Web Consortium:	http://www.w3c.org
Yahoo:	http://www.yahoo.com
Resources of General biochemical interest:	Table 14.9
Primary DB of biochemical interest:	Table 14.10
Gene Ontology:	Table 14.11

CHAPTER 15

GENOMICS

15.1 GENOME: FEATURES AND ORGANIZATION

15.1.1 Genome features

Genome is a collective term for the genetic material of an organism. It refers to the entire DNA content of a cell, including the nucleotide, genes and chromosomes (Singer and Berg, 1991). The field of study focused on the computational analysis of genomes is known as genomics (Benfey and Protopapas, 2005; Canter and Smith, 1999; Starkey and Elaswarapu, 2001). Genomes of living organisms have a profound diversity (Singer and Berg, 1991; Miklos and Rubin, 1996). This diversity relates not only to genome size but also to the storage principle as either single- or double-stranded DNA or RNA. Moreover, some genomes are linear (e.g. mammals) whereas others are closed and circular (e.g. most bacteria). Cellular genomes are always made of DNA, while phage and viral genomes may consist of either DNA or RNA. In single-stranded genomes, the information is read in the positive sense, the negative sense or in both directions, in which case we speak of an ambisense genome. The positive direction is defined as going from the 5′ to the 3′ end of the molecule. In double-stranded genomes, the information is read only in the positive direction (5′ to 3′ on either strand). Such is the diversity of the genome structure that even genomes of organisms that belong to the same taxonomic class often exhibit great diversity (genome diversity).

The smallest genomes are found in non-self-replicating bacteriophages and viruses. The very small genomes normally occur as one continuous piece of sequence. But other larger genomes may have several chromosomal components. For example, the approximately 3 billion (giga)-base pairs (Gbp) human genome, with approximately 30 000 protein-coding genes (Claverie, 2001), is organized into 22 chromosomes plus the two (X and Y chromosomes) that determine sex. The DNA contents of the autosomes range between 48 to 279 million (mega)-base pairs (Mbp). The X and Y chromosomes contain 163 and 51 Mbp respectively. Viral genome have sizes in the interval from 3.5 to 280 kilobase pairs (Kbp), bacteria range from 0.5 to 10 Mbp, fungi range from around 10–50 Mbp, plants start at around 50 Mbp, and mammals are found to be around 1–10 Gbp (Table 15.1)

In 1995, the first complete genome of a free-living organism, the prokaryote *Haemophilus influenzae*, was published and made available for analysis (Fleischmann *et al.*, 1995). This circular genome contains 1 830 137 bp with 1743 predicted protein coding regions and 76 genes encoding RNA molecules. The human genome consisting of 2.91 billion bp of DNA has been completely sequenced (Venter *et al.*, 2001) and the sequence can be obtained from GenBank (http://www.ncbi.nlm.nihgov/GenBank). Although the number of genes present has been estimated to be only about 30 000 (i.e. ~1%), it should be remembered that they are transcribed to pre-mRNA and several mechanisms serve to increase the variability of the expressed genes. The majority of genes are

---

Biomacromolecules, by C. Stan Tsai
Copyright © 2007 John Wiley & Sons, Inc.

TABLE 15.1 Size of some DNA molecules

Organism	Number of base pairs (bp)	Contour length (μm)	Max. possible number of proteins encoded
Viruses:			
Polyoma, SV40	5.1×10^3	1.7	4.25
Bacteriophage λ	4.86×10^4	17	40.5
Bacteriophages, T2/T4/T6	1.66×10^5	55	138
Bacteria:			
Mycoplama hominis	7.60×10^5	260	633
Haemophilus influenzae	1.86×10^6	632	1.55×10^3
Eschericia coli	4.70×10^6	1.60×10^3	3.917×10^3
Yeast (in 17 haploid chromosomes)	13.5×10^6	4.60×10^3	11.25×10^3
Drosophila (in 4 haploid chromosomes)	16.5×10^7	56×10^3	13.75×10^4
Human (in 23 haploid chromosomes)	2.90×10^9	990×10^3	2.417×10^6
Lungfish (in 19 haploid chromosomes)	10.2×10^{10}	34.7×10^6	8.50×10^7

Notes: 1. Assuming 1200 bp per protein.
2. Haploid refers to a single set of chromosomes.

subjected to alternative splicing after transcription, resulting in several gene products from each gene (Johnson *et al.*, 2003). Other RNA modifications as well as posttranslational modifications of proteins add complexity to the expression of the genome. Generalizations for DNA as the genome of many different organisms can be summarized as:

1. Prokaryotic genes are continuous regions of DNA. The functional unit of genetic sequence information from bacterium is a string of 3N nucleotides encoding protein with a string of N amino acid residues or N nucleotides encoding structural RNA of N residues. Eukaryotic genes appear split into separated segments in the genomic DNA. An exon is a stretch of DNA retained in the mature mRNA that is translated into protein. An intron is an intervening region between two exons.
2. The bacterial genome (e.g. *E. coli*) has little wasted space. Many of the protein-coding genes are arranged in operons, which when transcribed usually result in polycistronic mRNA. Apparently little nonfunctional DNA is contained both within and between operons (clusters of genes coordinately regulated at single promoter sites). The only repetitive DNAs are for ribosomal genes, insertion sequences and transposons.
3. Eukaryotic single cell organisms (e.g. yeast) appear to have little extra DNA aside from transposons and the relatively few introns (intervening sequences, regions removed from functional RNA products).
4. All higher organisms usually have linear chromosomal DNA molecules, although circles can be produced under special circumstances.
5. The protein-coding portion of eukaryotic cells of metazoans and higher plants make up a small fraction (~1%) of the total DNA of the cell. Distinct functional transcription units (the DNA regions that are transcribed to produce the primary transcripts) are most often separated by great distance.
6. The extra DNA in eukaryotes consists of introns, pseudogenes, simple sequence DNA, transposable elements and unclassified spacer DNA between transcription units. Sequence analysis of pseudogenes has shown that these genes were formerly

normal but has lost critical nucleotide sequences. Introns are present in some transcription units from all eukaryotic cell types, and they are abundant in vertebrate genes where they may make up to 80–90% of all DNA contained within transcription units.

7. DNA rearrangement may play an important role in gene expression during differentiation toward greater cell specialization.
8. Four general types of genomic DNA rearrangements in chromosome have been found, namely insertion, transposition, amplification and deletion. Insertion and transposition involve the movement of mobile genetic elements, insertion sequences (ISs) and transpositions respectively into a chromosomal site. The DNA of all cells contains elements that can move from site to site in the genome. In bacteria these elements move a DNA-copying mechanism that occurs simultaneously with the transposition. Two ISs can bracket a segment of DNA to form a transposon, which can then move in the genome. Eukaryotic transposons seem to move in the genome via an RNA intermediate. Specific chromosomal segments can undergo amplification or repeated localized copying. The deletion of specific sequences brings together parts of a transcriptional unit and thus activates genes.
9. Meiotic recombination is the breakage and reunion of DNA molecules prior to their segregation to daughter cells during the process of gamete formation. In the human genome, meiotic recombination occurs at a frequency of about 1% for two loci spaced 1 Mb apart. The actual frequency observed between two loci is called θ. We define a 1% recombination frequency as 1 cM (centiMorgan). Most human chromosomes are more than 100 cM in length.
10. A site of a gene on a chromosome, i.e. site on a DNA molecule is called the locus and a variant at a particular locus; namely a DNA sequence variant at the particular locus of interest, is termed an allele. What we observe, namely the complete set of observable inherited characteristics of an organism, is known as the phenotype, while what we obtain, namely the genetic constitution, which is the particular set of alleles inherited by the organism as a whole, is known as the genotype. Haplotype then refers to the actual pattern of alleles on one chromosome, i.e. on a single DNA molecule.

Table 15.2 lists some useful genome databases for general and related topics.

Concerning the genome organization, each of the organisms exhibits their own specific and often unexpected characteristics, for example an interesting observation concerning the distribution/amount of the genome encoded for genes (gene density). In *E. coli* and *S. Cerevisiae*, the genes are more or less equally distributed on both strands, whereas in *B. subtilis*, a strong bias in the polarity of transcription of the genes with respect to the replication fork is observed. Furthermore, although insertion sequences are widely distributed in bacteria, none were found in *B. subtilis*. For *A. thaliana*, an unexpectedly high percentage of protein shows no significant homology to proteins of organisms outside the plant kingdom. In prokaryotic organisms, i.e. eubacteria and archaea, the genome sizes vary considerably. However, their gene density is relatively constant at about one gene per kbp. During the evolution of eukaryotic organisms, genome size grew, but gene density decreased from one gene in 2 kbp in *Saccharomyces* to one gene in 10 kbp in *Drosophila*. Comparisons of protein sequences have shown that certain gene products can be found in a wide variety of organisms. Genomes with low gene density can cause problems in the accurate identification of genes.

TABLE 15.2 Useful general genome databases

Database	URL	Description
COGENT	http://maine.ebi.ac.uk:8000/service/cognet/	Complete genome tracking
EBI Genomes	http://www.ebi.ac.uk/genomes	EBI C/U genomes
EMGlib	http://pbil.univ-lycon1.fr/emglib/emglib.html	Unicellular organism genomes
Entrez Genomes	http://www.ncbi.nlm.nih.gov/entrez/	NCBI C/U genomes
Genew	http://www.gene.ucl.ac.uk/nomenclature	Human gene nomenclature
GeneLetter	http://www.geneletter.org/	GP news
Gene3D	http://www.biochem.ucl.ac.uk/bsm/cath_new/Gene3D/	Structural assignments for whole genomes
Genome Atlas	http://cbs.dtu.dk/services/GenomeAtlas/	DNA structural properties of sequenced genomes
Genome Info Broker	http://gib.genes.nig.ac.jp	DDBJ C/U genomes
Genome Reviews	http://www.ebi.ac.uk/GenomeReviews/	Integrated view of complete genomes
GenomeWeb	http://www.hgmp.mrc.ac.uk/GenomeWeb/	GP general information
GO	http://www.geneontology.org/	Gene ontology Consortium
GOLD	http://www.genomesonline.org/	Listing of C/O GP, links
MBGD	http://mbgd.genome.ad.jp/	Comparative microbial genomes
NCBI GEO	http://www.ncbi.nlm.nih.gov/GEO	Gene expression data
NCBI Taxonomy	http://www.ncbi.nlm.nih.gov/Taxonomy/	All organisms in GenBank
ORNL	http://www.ornl.gov/TechResources/Human_Genome/home.html	Human GP information
PANTHER	http://panther.celera.com/	Gene products by biol. functions
PEDANT	http://pedant.gsf.de/	Analysis of genomic sequences
TransportDB	http://www.membranetransport.org/	Membrane transporter in completed genomes
TIGR	http://tigr.org/tdb/mdb/mdbcomplete.html	Lists of C/O GP, links to sequences

Note: Abbreviations used: C/O, completed and ongoing; C/U, completed and unfinished; GP, genome projects; NR, non-redundant.

Comparative genomics compares all the gene sequences of a particular organism with all other genomes, in order to identify differences that may account for defined and important properties. Structural genomics aims to expedite the determination of protein structures and relate structures (Bourne and Weissig, 2003) to gene sequences and henceforth to functions in functional genomics (Pevsner, 2003).

15.1.2 Gene mapping

There are two types of gene mapping of chromosomes, i.e. physical mapping and genetic mapping. An ordered set of gene loci with relative spacing (and order) determined from measured recombination frequencies is called a genetic map, while a physical map is based directly on measurements of DNA structure. Physical mapping uses a variety of methods to assign genes and DNA markers to particular locations along a chromosome, so the actual distances between the genes (measured in nucleotide bps) are known. Genetic mapping describes the arrangement of genes based on the relationship of their linkage. DNA markers or probes can also be used in the construction of genetic maps if they detect sequence changes (polymorphism) among different individuals. The tendency of two genes or DNA markers to segregate together through meiosis in family studies gives a description of genetic linkage, but not their physical location. The order of genes on a chromosome measured by genetic linkage is the same as the order in physical maps, but there is no constant scale factor that relates physical and genetic distance. The variation in scale occurs because recombination does not occur at equal frequencies for different intervals along a chromosome. Since most of the genes and DNA markers used in the construction of genetic maps exist as cloned and sequenced DNA fragments, they can also be readily placed on a physical map. Genetic and physical maps for human, mouse and rat can be accessed from WIGS at http://www.genome.wi.mit.edu/

Four known gene maps of chromosomes are:

1. *Gene linkage maps* express recombination frequencies of linked genes on the chromosome. The unit of length in a gene map is the Morgan defined as 1 cM = 1% recombination frequency (in human, 1 cM ≈ 10^6 bp).

2. *Chromosome physical maps* express chromosomes by physical objects such as their banding patterns and their visible features. The banding patterns on chromosomes are numbered in order of size, with 1 being the largest. The two arms of human chromosomes, separated by the centromere, are called p (petite) arm and q (queue) arm, numbering p1, p2 . . . and q1, q2 . . . outward from the centromere with subsequent digits for subdivisions of bands (e.g. 15q11.2 for chromosome 15, sub-band 2 of band 11).

Physical maps are needed because ordinary human genetic maps are not detailed enough to allow the DNA that corresponds to particular genes to be isolated efficiently. Physical maps are also needed as the source for the DNA samples that can serve as the actual substrate for large-scale DNA sequencing project. The three major goals in physical mapping:

1. provide an ordered set of all the DNA of a chromosome (or genome).
2. provide accurate distance between a dense set of DNA markers.
3. provide a set of DNA samples from which direct DNA sequencing of the chromosome or genome is possible.

3. *Restriction maps*: A typical restriction map consists of an ordered set of DNA fragments that can be generated from a chromosome by cleavage with restriction enzymes individually or in pairs. Distances along the map are known precisely as the lengths of the DNA fragments generated by the enzymes can be measured. They are accurate to a single bp up to DNAs around 1 kb in sizes. Lengths can be measured with better than 1% accuracy for fragments up to 10 kb, and with a low percent of accuracy for fragments up to 1 Mb in size. A search for the restriction profile can be performed at NEBcutter (http://tools.neb.com/NEBcutter2/index.php) and Webcutter (http://www.firstmarket.com/cutter/cut2.html). Advantages of a restriction map are:

- Accurate lengths are known between sets of reference points.
- Most restriction mapping can be carried out using a top-down strategy that preserves an overview of the target and that reaches a nearly complete map relatively quickly.
- We are working with genomic DNA fragments rather than cloned DNA, thus all of the potential artifacts that can arise from cloning procedures are avoided.

In top-down mapping, we successively divide a chromosome target into finer regions and order these. In a typical restriction mapping effort, any pre-existing genetic map information can be used as a framework for constructing the physical map. Large DNA fragments are produced by cutting the chromosome with restriction enzymes with rare recognition sites. The fragments are separated by size and assigned to regions by hybridization with genetically or cytogenetically mapped DNA probes. The fragments are assembled into contiguous blocks to give a macrorestriction map. If a finer map is desired, it can be constructed by taking the ordered fragments one at a time and dissecting these with more frequently-cutting restriction nucleases.

4. *DNA sequences*: Genes are regions of the sequences of dexoyribonucleotides. Only with DNA sequences, the stored hereditary information in its molecular form is expressed. They are computationally represented by strings of characters A, T, G and C.

15.1.3 Information content of nucleotide sequence

Sequence segments of DNA, which encode protein products or RNA molecules, are called coding regions, whereas those segments that do not directly give rise to gene products are normally called noncoding regions. A coding sequence (CDS, cds) is a subsequence of a DNA sequence that is surmised to encode a gene. A CDS begins with an initiation codon and ends with a stop codon. In the cases of spliced genes, all exons and introns should be within the same CDS. Noncoding regions can be parts of genes, either as regulatory elements or as intervening sequences (IVSs, introns). Untranslated regions (UTRs) occur in both DNA and RNA. They are portions of the sequence flanking the final coding sequence. The 5′UTR at the 5′ end contains the promoter site and the 3′UTR at the 3′ end is highly specific both to the gene and to the species from which the sequence is derived.

It is possible to translate a piece of DNA sequence into protein by reading successive codons with reference to a genetic code table. This is termed the conceptual translation that has no biological validation or significance. Because it is not known whether the first base marks the start of the CDS, it is always essential to perform *a six frame translation*. This includes three forward frames that are accomplished by beginning to translate at the first, second and third bases respectively, and three reverse frames that are achieved by reversing the DNA sequence (the complementary strand) and again beginning on the first, second and third bases. Thus for any DNA genome, the result of a six frame translation is six potential protein sequences, of which only one is biologically functional.

Computer programs for genome analysis generally identify reading frames. A reading frame is a region of DNA sequence that begins with an initiation codon (ATG) and ends with the stop codon (TAA, TAG or TGA). The information for genetic codes and codon usage can be obtained from CUTG (http://www.kazusa.or.jp/codon/), Genetic Codes (http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi) and Transterm (http://uther.otago.ac.nz/Transterm.html). The correct reading frame is normally taken to be the longest frame uninterrupted by a stop codon. Such a frame is known as an open reading frame (ORF). An ORF corresponds to a stretch of DNA that could potentially be translated into a polypeptide. For an ORF to be considered as a good candidate for coding a cellular protein, a minimum size requirement is often set, e.g., a stretch of DNA that would code for a protein of at least 100 amino acids. An ORF is not usually considered equivalent to a gene until there has been shown to be a phenotype associated with a mutation in the ORF, and/or an mRNA transcript or a gene product generated from the DNA molecule of ORF has been detected.

The initial codon in the CDS is that for methionine (ATG), but methionine is also a common residue within the CDS. Therefore its presence is not an absolute indicator of ORF initiation. It is also noted that the stop codon TGA is translated into a selenocysteine residue in selenoproteins. Several features may be used as indicators of potential protein coding regions in DNA such as sufficient ORF length, flanking Kozak sequence (CCGCCATGG) (Kozak, 1996), species specific codon usage bias and detection of ribosome binding sites upstream of the start codon of prokaryotic genes. Ultimately the surest way of predicting a gene is by alignment with a homologous protein sequence. A DNA segment with repeat sequences is an indication of an unlikely protein coding region whereas a segment with apparent codon bias is an indicator of the protein coding region. Sequence similarity to other genes or gene products provides strong evidence for exons and matches to template patterns may indicate the locations of functional sites on the DNA.

The eukaryotic genes are characterized by:

- regions that contribute to the CDS, known as exons; and
- those that do not contribute, known as introns.

The presence of exons and introns in eukaryotic genes results in potential gene products with different lengths because not all exons are jointed in the final transcribed mRNA. The proteins resulting from the mRNA editing process with different translated polypeptide chains are known as splice variants or alternatively spliced forms. Therefore database searches showing the sequences with substantial deletions in matches to the query sequences could be the result of alternative splicing. On-line bibliographies of publications relevant to analysis of nucleotide sequences are maintained at SEQANALREF (http://expasy.hcuge.ch).

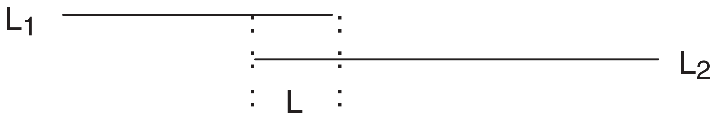
15.1.4 DNA library

In order to clone a gene, we begins by constructing a DNA library. Cleaving the entire genome of a cell with a specific restriction endonuclease is called the whole genome shotgun (WGS) approach to gene cloning. This strategy involves shearing the entire DNA of an organism into segments of a defined length, which are cloned into a plasmid vector for DNA sequencing (Weber and Myers, 1997). Such a plasmid is said to contain a genomic DNA clone and the entire collection of plasmids is said to comprise a genomic library. Sufficient DNA sequencing is performed so that each nucleotide of DNA in the genome is covered numerous times in fragments of about 500 bp. Large scaffolds of DNA sequences can be assembled by identifying overlapping stretches of DNA sequences. These can then be ordered, extended and assembled to the complete genome.

In genomics, we often refer to clone libraries (ordered libraries). Most genomic libraries are made by partial digestion with relatively frequent-cutting restriction enzymes, size selection of the fragments to provide a fairly uniform set of DNA inserts, and cloning these into a vector appropriate for the size range of interest. The cloned fragments are a nearly random set of DNA pieces. Because the clones contain overlapping regions of the genome, it is possible to detect these overlaps by various fingerprinting methods that examine patterns of sequence on particular clones. The random nature of the cloned fragments means that many more clones exist than the minimum set necessary to cover the genome. In practice, the redundancy of the library is usually set at five- to tenfold in order to ensure that almost all regions of the genome will have been sampled at least once. From this vast library the goal is to assemble and to order the minimum set of clones that covers the genome in one contiguous block. This set is called the tiling path.

Clone libraries have usually been ordered by a bottom-up approach. Here individual clones are initially selected from the library at random. Usually the library is handled as an array of samples so that each clone has a unique location on a set of microtitre plates. The clone is fingerprinted, by hybridization, by restriction mapping or by determining bits of DNA sequence. Eventually clones overlap, if not identical, and they are assembled into overlapping sets called contigs (contiguous blocks), which is a series of overlapping DNA clones of known order along a chromosome from an organism of interest. While this process is easy to automate, the fingerprinting and contig building provide no overview of chromosome or genome. Additional experiments have to be done to place contigs on a lower-resolution framework map. Usually the overlap distance between two adjacent members of a contig is not known with much precision. Therefore distances on a typical bottom-up map are not well defined.

Thus the conventional approach to constructing an ordered bottom-up library is to fingerprint a dense set of samples and look for clones that share overlapping properties. A key variable in designing such strategies is the minimum fraction of the clones that must be in common in order for their overlap to be detectable. The smaller the overlap required, the fewer clones needed to produce an ordered set and the faster the process proceeds. The degree of overlap, f (where $f = L/[L_1 + L_2 - L]$) of two clones of length L_1 and L_2 will determine the resolution of the fingerprinting procedure needed to identify them.



Typical methods usually require a five- to tenfold redundant set of clones to ensure that there is a good chance of sampling the entire target at sufficient density to allow overlap detection.

An approach for clone fingerprinting is by the use of restriction enzyme digestion. Indirect end labeling from probes in the vector sequence is used to determine the positions of restriction sites seen in separate partial digests, each generated with one of several different restriction enzymes. The digests are analyzed in a single gel electrophoresis. Thus the relative order of the different restriction sites would be known with great accuracy. The order information is the major source of data used to fingerprint the clones. Two features of DNA in human genomes that vary among individuals can serve as intrinsic genetic markers. Variable number tandem repeats (VNTRs) or minisatellites contain regions 10–100 bp long with the same sequence repeated various number of times with different lengths on different chromosomes. Inheritance of VNTR can be followed in a family for personal identification, i.e. genetic fingerprints. Short tandem repeat polymorphisms (STRPs) or microsatellites are regions of 2–5 bp that repeat many times (typically 10–30 consecutive copies) and are evenly distributed over the human genome.

To prepare cDNA library (Cowell and Austin, 1997) suitable for use in rapid sequence experiments, a sample of cells is obtained, then mRNA is extracted from the cells and is employed as the template for reverse transcriptase to synthesize cDNA, which is transformed into a library. A sample of clones each between 200 and 500 bases representing part of the genome is selected from the library at random for sequence analysis. The sequences that emerge from this process are called expressed sequence tags (ESTs). Thus EST is a specialized complementary DNA molecule that has been subjected to a single pass of DNA sequencing. Good libraries contain at least 1 million clones and probably substantially more though the actual number of distinct genes expressed in a cell may be a few thousand. The number varies according to cell type (e.g. human brain cells with ~15 000 genes and gut cells with ~2000 genes). Thus a large part of currently available DNA data is made up of partial sequences, the majority of which are ESTs. The following considerations must be taken into account in analyzing ESTs:

- There are five characters, ACGTN (N for any base), in the EST alphabet.
- There may be phantom InDels, resulting in translation frameshifts.
- The EST will often be a sub-sequence of any other sequence in the database.
- The EST may not represent part of the CDS of any gene.

The EST databases probably contain fragments of a majority of all genes. Thus they are an important resource for locating some part of most genes. Some of the EST databases are given in Table 15.3.

15.1.5 Alternative splicing

Alternative splicing leads to proteome expansion bridging a perceived complexity gap in human genome. The role of alternative splicing (also known as functional splicing) in generating proteome diversity is augmented by a role in regulation. As a regulatory process, alternative splicing contributes to biological complexity through its ability to control the expression of proteins. Table 15.4 provides sample databases for alternative splicing.

TABLE 15.3 Some EST databases

Database	URL	Description
dbEST	http://www.ncbi.nlm.nih.gov/dbEST/index.html	EST repository
TIGR Gene Indices	http://www.tigr.org/tdb/tgi.shtml	Organism specific EST and gene sequences
UniGene	http://www.ncbi.nlm.nih.gov/UniGene/	Unified clusters of EST and full-length mRNA sequences
UniSTS	http://www.ncbi.nlm.nih.gov/entrez/query.fcg?db=unists	Unified view of sequence tagged sites with mapping data
GSC, WUSTL	http://genome.wustl.edu/gsc/	Human and model organism sequences, EST
EASED	http://eased.bioinf.mdc-berlin.de/	Extended alternatively spliced EST
Gene Resour. Locator	http://grl.gi.k.u-tokyo.ac.jp/	Alignment of EST with finished human sequences
ApiEST-DB	http://www.cbil.upenn.edu/paradbs-sevlet	*Apiconplexan* parasites EST seq
COGEME	http://cogeme.ex.ac.uk	Phytopathogenic fungi and oomycete EST
CR-EST	http://pgrc.ipk-gatersleben.de/cr-est/	Crop EST
Diatom EST DB	http://avesthagen.sznbowler.com/	EST from diatom algae
LumbriBase	http://www.earthworms.org/	EST of *Lumbricus rubellus*
Mendel	http://www.mendel.ac.uk/	Annotated plant EST
openSputnik	http://sputnik.btk.fi	Functional annotation of plant EST
PEDE	http://pede.gene.staff.or.jp/	Pig EST and cDNA
SilkDB	http://www.ab.a.u-tokyo.ac.jp/genome/	Silkworm EST

TABLE 15.4 Databases for alternative splicing

Database	URL	Description
ASAP	http://www.bioinformatics.ucla.edu/ASAP	AS isoforms
ASD	http://www.ebi.ac.uk/asd	Exons from different sources: AltSlice, AltExtron, Aedb
ASDB	http://hazelton.lbl.gov/~teplitsk/alt	Protein products of AS genes
EASED	http://eased.bioinf.mdc-berlin.de/	EST inferred AS
ECgene	http://genome.ewha.ac.kr/ECgene/	Genome annotation for AS
EDAS	http://www.ig-msk.ru:8005/EDAS/	EST-derived AS DB
ASHESdb	http://sege.ntu.edu.sg/wester/ashes/	AS human genes by exon skipping DB
Intronerator	http://www.cse.ucsc.edu/~kent/intronerator/	AS in *C. elegans, C. briggsae*
MAASE	http://splice.sdsc.edu/	AS search
PASDB	http://pasdb.genomics.org.cn/	AS literature

Note: Abbreviations used: AS, alternative(ly) spliced; DB, database.

15.1.6 Gene variation: Single nucleotide polymorphism

Single nucleotide polymorphism (SNP) is a genetic variation between individuals, limited to a single bp, which can be substituted, inserted or deleted. Changes in the primary sequences of genes and proteins can lead to disease phenotypes. For example, sickle-cell anemia is a disease caused by a specific SNP, an A→T mutation in the β-globin gene resulting in a Glu-6→Val-6 change in β subunit of hemoglobin (Hb A versus Hb S). SNP is distributed throughout the genome (on the average every 2 kbp). Although they arose by mutation, many positions containing SNP have low mutation rates and provide stable markers for mapping genes. Common methods used to scan genes for mutations (Strachan and Read, 1999) are given in Table 15.5. Some SNPs that occur within exons are mutations to synonymous codons, or cause substitutions that do not significantly affect protein function. Other types of SNPs can cause more than local perturbation to a protein, such as:

- a mutation from a sense codon to a stop codon or *vice versa* will cause either premature truncation of protein synthesis or readthrough; and
- a deletion or insertion will cause a phase shift in translation.

Strong correlation of a disease with a specific SNP is advantageous in clinical work because it is relatively easy to test for the affected individuals or carriers. A particular site may predominate if:

- all bearers of the gene are descendants of a single individual in whom the mutation occurred;

TABLE 15.5 Common methods used to scan genes for mutations

Method	Advantages	Disadvantages
Southern blot (cDNA probe)	Detect major deletions/rearrangements	Labor intensive and requires μg of DNA
Sequencing	Detect all changes and mutations are fully characterized	Expensive/time-consuming and interpretation may be difficult
Heteroduplex gel mobility	Simple, cheap	Sequences of <200 bp only, limited sensitivity and does not reveal position of the change
Denaturing HPLC	Quick, high throughput and quantitative	Expensive and does not reveal position of change
Single-strand conformation polymophism (SSCP) analysis	Simple, cheap	Sequences of <200 bp only, limited sensitivity and does not reveal position of the change
Denaturing gradient gel electrophoresis	Highly sensitive	Expensive, can be technically difficult and does not reveal position of change
Dideoxy fingerprinting	Highly sensitive	Interpretation may be difficult
Mismatch cleavage, chemical or enzymatic	Highly sensitive, indicates position of change	Experimental difficulty
Protein truncation test (PTT)	Highly sensitive for chain terminating mutations, shows the position of change	Detects mutations that lead to protein truncation only, expensive and technically difficult
Microarray analysis	Quick and high throughput, able to potentially detect and define all changes	Expensive and restricted to a predefined number of genes

- the disease results from a gain rather than loss of a specific property; and/or
- the mutation rate at a particular site is unusually high.

Other clinical applications of SNP reflect correlation between phenotype and reaction to therapy (pharmacogenomics). SNP is the basis for ushering in the personalized medicine. For example, an SNP in the gene for *N*-acetyl transferase is correlated with peripheral neuropathy (weakness, numbness and pain in the arms, legs, hands and feet) as the side effect of treatment for tuberculosis with isoniazid (isonicotinic acid hydrazide). Therefore patients with this SNP are given alternative treatment.

DNA microarrays (chips) offer a high-throughput technology for monitoring mutations (Table 15.5) in human disease genes, for mapping disease genes by linkage analysis and for discovering novel SNP. To expedite microarray analysis of sequence variants, it is useful to have informatics tools that allow the search and retrieval of such sequences. One straightforward approach is to use PubMed and the scientific literature to identify mutations and other sequence variants described in experimental studies. Another approach is to use SNP databases (Table 15.6) such as dbSNP provided by NCBI (www.ncbi.nlm.nih.gov/SNP/) and SNP Consortium (http://snp.cshl.org). The dbSNP contains the sequences of SNPs, small insertion and deletion mutations, and microsatellite repeats and therefore provides access to a broad spectrum of minor sequence variants. The dbSNP can be queried using a keyword of interest such as ‘beta-globin’, and the results of searches provide the location of mutations within each database sequence, listed as a function of the GenBank accession number. Electronic access to sequence variant databases expedites disease testing, screening, diagnostic and other applications of microarrays that focus on DNA information.

15.2 GENOME INFORMATICS: DATABASES AND WEB SERVERS

15.2.1 Nucleic acid databases

Nucleic acid sequence databases typically contain sequence data, which includes information at the level of the gene structures, introns and exons (for eukaryotics), cDNA (complementary DNA), RNA and transcription regulations. The important nucleic acid sequence data repositories as the primary resources known as International Nucleotide Sequence Database Collaboration (INSDC) are:

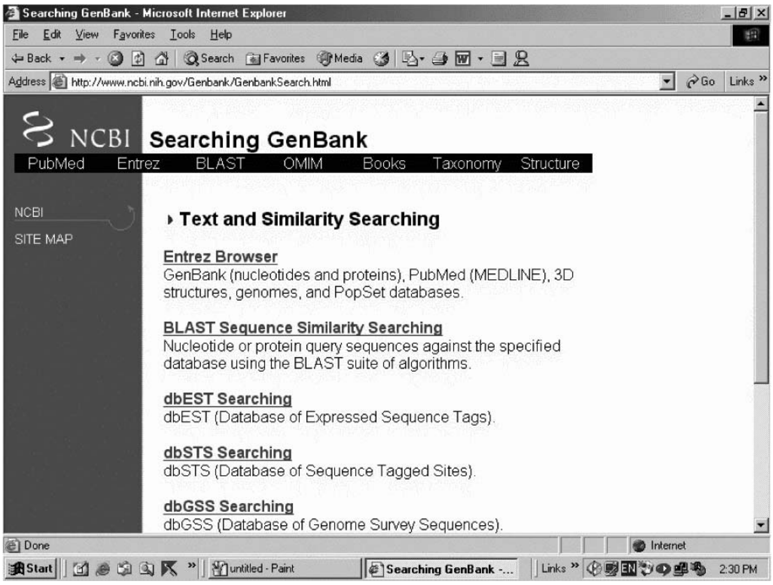
TABLE 15.6 Some SNP databases

Database	URL	Description
dbQSNP	http://qsnp.gen.kyushu-u.ac.jp	Quantification SNP allele frequencies
dbSNP	http://www.ncbi.nlm.nih.gov/SNP/	NCBI SNP
FESD	http://combio.kribb.re.kr/ksnp/resd/	Functional elements of SNP
JSNP	http://snp.ims.u-tokyo.ac.jp/	SNP
rSNP	http://wwwmgs.bionet.nsc.ru/mgs/systems/rsnp/	SNP in regulatory gene regions
SNP Consortium	http://snp.cshl.org/	SNP Consortium
SNPeffect	http://snpeffect.vib.be/	Phenotypic effects of human SNP
TopoSNP	http://gila.bioengr.uic.edu/snp/toposnp	Topography of nonsynonymous SNP
WICGS	http://www-genome.wi.mit.edu/	Human SNP

GenBank of the NCBI:	http://www.ncbi.nlm.nih.gov/Genbank/
EMBL Nucleotide Sequence DB at EBI:	http://www.ebi.ac.uk/embl.html
DNA Data Bank of Japan (DDBJ):	http://ddbj.nig.ac.jp

All three centers are separate points of data submission, but they all exchange this information and make the same database available to the international communities.

GenBank (Benson *et al.*, 2006), the DNA database from the National Center for Biotechnology Information (NCBI), incorporates sequences from publicly available sources, primarily from direct author submissions and genome sequence projects. The resource exchanges data with EMBL and DDBJ on a daily basis to ensure comprehensive coverage worldwide. GenBank is accessed at http://www.ncbi.nlm.nih.gov/GenBank (Figure 15.1) or via Entrez (http://www.ncbi.nlm.nih.gov/Entrez). To facilitate fast and specific searches, the GenBank database provides Entrez browser, similarity search of nucleotide and amino acid sequence. Searches for expressed sequence tag (EST), sequence tagged site (STS) and genome survey sequence (GSS) databases are also available from the GenBank. Each GenBank entry consists of a number of keywords. The LOCUS keyword introduces a short label and other relevant facts including the number of base pairs, source of sequence data, division of database and date of submission. A concise description of the sequence is given after the DEFINITION and a unique accession number after the ACCESSION. The KEYWORDS line includes short phrases assigned by author, describing gene products and other relevant information. The SOURCE keyword and ORGANISM sub-keyword indicate the biological origin of the entry. The REFERENCE and its sub-keywords record bibliographic citation with the MEDLINE line pointing to an online link for viewing the abstract of the given article. The FEATURES keyword introduces the feature table describing properties of the sequence in detail and relevant cross-linked information. The BASE COUNT line provides the frequency count of different


Searching GenBank - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.ncbi.nih.gov/Genbank/GenbankSearch.html
Go Links
NCBI
PubMed Entrez BLAST OMIM Books Taxonomy Structure
NCBI
SITE MAP
Done
Internet
Start
untitled - Paint
Searching GenBank -...
Links
2:30 PM

Searching GenBank

Text and Similarity Searching

Entrez Browser
GenBank (nucleotides and proteins), PubMed (MEDLINE), 3D structures, genomes, and PopSet databases.

BLAST Sequence Similarity Searching
Nucleotide or protein query sequences against the specified database using the BLAST suite of algorithms.

dbEST Searching
dbEST (Database of Expressed Sequence Tags).

dbSTS Searching
dbSTS (Database of Sequence Tagged Sites).

dbGSS Searching
dbGSS (Database of Genome Survey Sequences).

Figure 15.1 Search page of GeneBank. The search facility of GenBank (http://www.ncbi.nlm.nih.gov/GenBank/GenbankSearch.html) provides access to Entrez browser and various searching engines for BLAST similarity for nucleotide and amino acid sequences, dbEST, dbSTS and dbGSS

bases in the sequence. The ORIGIN line records the location of the first base of the sequence within the genome if known. The nucleotide sequence of the genome (in GenBank format, Chapter 4) follows, and the entry is terminated by the // marker.

EMBL (Kanz *et al.*, 2005), the nucleotide sequence database maintained by the European Bioinformatics Institute (EBI) (Emmert *et al.*, 1994) at http://www.ebi.ac.uk/, produces sequences from direct author submissions and genome sequencing groups, and from the scientific literature and patent applications. The database is produced in collaboration with GenBank and DDJB. All new and updated entries are exchanged between the groups. Information can be retrieved from EBI using the SRS Retrieval System (http://srs.ebi.ac.uk/). In addition to sequence retrieval, EBI offers a wide variety of bioinformatic tools and services (Figure 15.2). The Institute has developed into one of the major genomics and proteomics resource/service sites. Each EMBL/EBI entry consists of the following lines with headings such as ID (identifier), AC (accession number), DE (description), KW (keyword/name), OS/OC (biological origin), RX/RA/RT/RL (reference pointer to MEDLINE, authors, title and literature), DR (pointer to amino acid sequence), FT (features) and SQ (the nucleotide sequence in Embl format is preceded by the base count and terminated with the // marker (Chapter 4).

DDBJ (Tateno *et al.*, 2005) is the DNA Data Bank of Japan in collaboration with GenBank and EMBL. The database is produced, maintained and distributed at the National Institute of Genetics. The entry of DDBJ follows the keywords adapted by the GenBank. DDBJ is accessed at http://www.ddbj.nig.ac.jp/ (Figure15.3).

Ribonucleic acids play vital roles in the flow of genetic information. A large number of RNA sequence databases, as given in Table 15.7, illustrate different types of RNA and their functional diversities.

The three-dimensional structures of oligonucleotides, DNA and RNA can be obtained from either NDB at http://ndbserver.rutgers.edu/ or PDB at http://

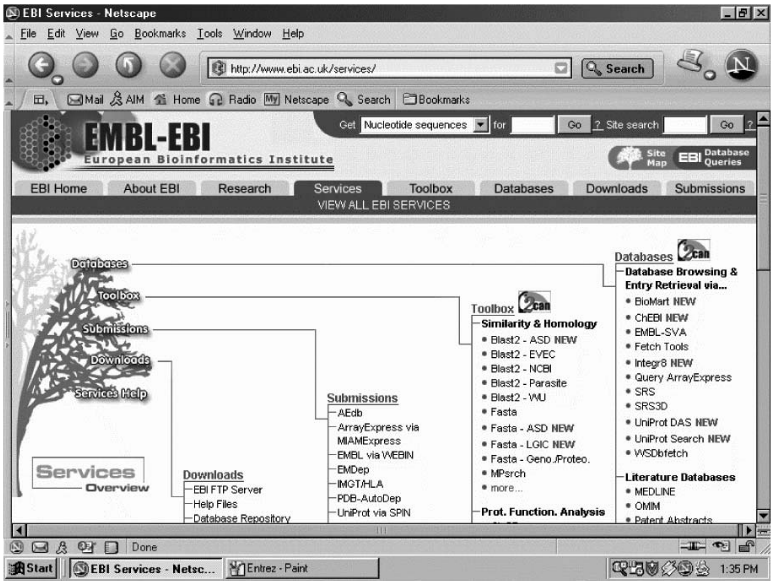
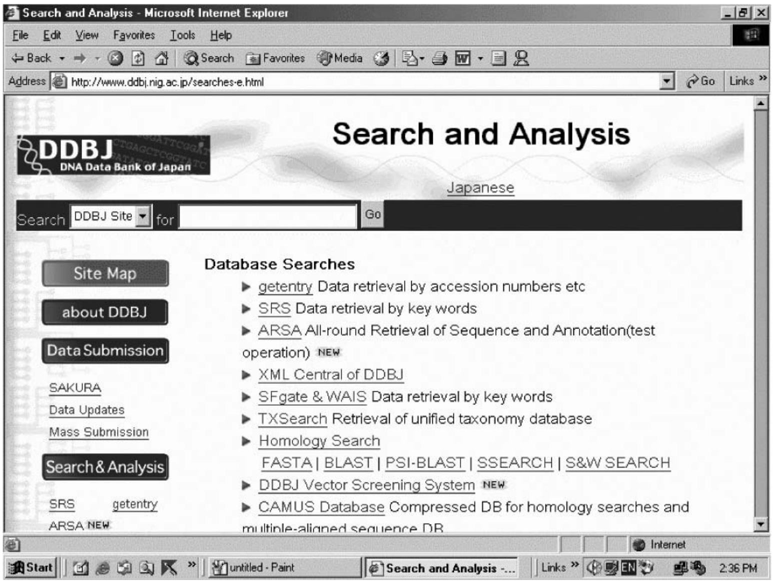

EBI Services - Netscape
File Edit View Go Bookmarks Tools Window Help
http://www.ebi.ac.uk/services/
Search
Mail AIM Home Radio Netscape Search Bookmarks
EMBL-EBI
European Bioinformatics Institute
Get Nucleotide sequences for Go ? Site search Go ?
Site Map
EBI Database Queries
EBI Home
About EBI
Research
Services
Toolbox
Databases
Downloads
Submissions
VIEW ALL EBI SERVICES
Databases
Toolbox
Submissions
Downloads
Services Help
Services Overview
Downloads
EBI FTP Server
Help Files
Database Repository
Submissions
AEdb
ArrayExpress via MIAMExpress
EMBL via WEBIN
EMDep
IMGT/HLA
PDB-AutoDep
UniProt via SPIN
Toolbox
Similarity & Homology
Blast2 - ASD NEW
Blast2 - EVEC
Blast2 - NCBI
Blast2 - Parasite
Blast2 - WU
Fasta
Fasta - ASD NEW
Fasta - LGIC NEW
Fasta - Geno./Proteo.
MPsrch
more...
Prot. Function. Analysis
Databases
Database Browsing & Entry Retrieval via...
BioMart NEW
ChEBI NEW
EMBL-SVA
Fetch Tools
Integr8 NEW
Query ArrayExpress
SRS
SRS3D
UniProt DAS NEW
UniProt Search NEW
WSDbfetch
Literature Databases
MEDLINE
OMIM
Patent Abstracts
Done
Start
EBI Services - Netsc...
Entrez - Paint
1:35 PM

Figure 15.2 Home page for services provided by EBI. Sequence and structure databases, analyses as well as various bioinformatic tools are accessible from the service page of EBI at http://www.ebi.ac.uk/services


Search and Analysis - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.ddbj.nig.ac.jp/searches-e.html
Go Links
Internet
Start
untitled - Paint
Search and Analysis -...
Links
2:36 PM

DDBJ DNA Data Bank of Japan

Search and Analysis

Japanese

Search DDBJ Site for Go

Site Map
about DDBJ
Data Submission
SAKURA
Data Updates
Mass Submission
Search & Analysis
SRS getentry
ARSA NEW

Database Searches
- getentry Data retrieval by accession numbers etc
- SRS Data retrieval by key words
- ARSA All-round Retrieval of Sequence and Annotation(test operation) NEW
- XML Central of DDBJ
- SFgate & WAIS Data retrieval by key words
- TXSearch Retrieval of unified taxonomy database
- Homology Search
  FASTA | BLAST | PSI-BLAST | SSEARCH | S&W SEARCH
- DDBJ Vector Screening System NEW
- CAMUS Database Compressed DB for homology searches and multiple-aligned sequence DB

Figure 15.3 Service and analysis page of DDBJ. The search and analysis facilities of DDBJ are accessible at http://www.ddbj.nig.ac.jp/searches-e.html. Data retrieval employs SRS

www.rcsb.org/pdb. On the NDB home page, choose Search then click the Execute Selection button to open a list of the database classes (DNA, DNA/RNA, Peptide nucleic acid, Peptide nucleic acid/DNA, Protein/DNA, Protein/RNA, Ribosome, Ribozyme, RNA and tRNA). Highlight an entry and click the Display Selection button to open an information page providing NDB ID, Compound name, Sequence, Citation, Crystal information, Coordinates and Views. The file can be saved (e.g. as .pdb file) after clicking the link, ‘coordinates for the asymmetric unit’ to open the coordinate file. RNABase (http://www.rnabase.org/) is an annotated database of RNA structures extracted from NDB and PDB (Murthy and Rose, 2003). Structural classification of RNA can be accessed at SCOR (http://scor.lbl.gov/) (Tamura *et al.*, 2004).

The DNA secondary databases offer analytical results (e.g. gene motifs, splice sties, transcription regulators) derived form the primary databases of INSDC, some of which are listed in Table 15.8.

15.2.2 Nucleic acid analysis servers

All primary sequence databases provide tools for essential sequence analyses. Many servers are also available on the web to perform useful computation on DNA/RNA sequences and structures. These web servers (Table 15.9) provide an array of diverse computational genomic tools.

15.3 APPROACHES TO GENE IDENTIFICATION

Genomics conducts structural and functional studies of genomes (McKusick, 1997; Shapiro and Harris, 2000; Starkey and Elaswarapu, 2001). The former deals with the determination of DNA sequences and gene mapping, while the latter is concerned with the attachment of functional information to existing structural knowledge about DNA sequences.

TABLE 15.7 Various RNA sequence databases

Database	URL	Description
5S rRNA	http://biobase.ibch.poznan.pl/5Sdata/	5S rRNA sequences
ARED	http://rc.kfshrc.edu.sa/ared	mRNA with AU-rich element
Mobile gr II introns	http://www.fp.ucalgary.ca/group2introns/	Group II intron ribozymes
Eur. RRNA	http://www.psb.ugent.be/rRNA/	rRNA sequences
GtRDB	http://rna.wustl.edu/GtRDB	Genomic tRNA
HIV Sequence DB	http://hiv-web.lanl.gov/	HIV RNA sequences
HuSiDa	http://itb1.biologie.hu-berlin.de/~nebulus/sirna	Human siRNA
HyPaLib	http://bibiserv.techfak.uni-bielefeld.de/HyPa/	Hybrid pattern library for RNA structural elements
IRESdb	http://ifr31w3.toulouse.inserm.fr/IRESdatabase/	Internal ribosome entry sites
miRNA Registry	http://www.sanger.ac.uk/Software/Rfam/mirna/	MicroRNAs
NCIR	http://prion.bchs.uh.edu/bp_type/	Non-canonical int. RNA structures
ncRNAs	http://biobases.ibch.poznan.pl/ncRNA/	NC RNA with regulatory functions
NONCODE	http://www.bioinfo.org.cn/NONCODE/index.htm	NC RNAs
PLANTncRNAs	http://www.prl.msu.edu/PLANTncRNAs	Plant NC RNAs
polyA_DB	http://polya.umdnj.edu/polyadb/	Mammal. mRNA polyadenylation
PseudoBase	http://wwwbio.leidenuniv.nl/~Batenburg/PKB.html	RNA peudoknots
RDP	http://rdp.cme.msu.edu/	rRNA sequences
Rfam	http://www.sanger.ac.uk/Software/Rfam/	ncRNA families
RISSC	http://ulises.umh.es/RISSC	Ribosomal internal spacer sequences
RNAdb	http://ncma.bioinformatics.com.au/	Mammalian NC RNA
RNA modification	http://medlib.med.utah.edu/RNAmods/	Modified nucleosides in RNA
RRNDB	http://rrndb.cme.msu.edu/	Prokaryotic rRNA operons
siRNAdb	http://mbcr.bcn.tmc.edu/smallRNA	Small interference RNAs
SRPDB	http://psyche.uthct.edu/dbs/SRPDB/SRPDB.html	Signal recognition particles
SSU rRNA Modif.	http://medlib.med.utah..edu/SSUmods/	Modified nucleosides in small subunit rRNA
Subviral RNA	http://subviral.med.uottawa.ca/	Viroid/viroid-like RNAs
tmRNA	http://www.indiana.edu/~tmrna	tmRNA sequences/alignments
tmRDB	http://psyche.uthct.edu/dbs/tmRDB/tmRDB.html	tmRNA
tRNA sequences	http://www.uni-bayreuth.de/departments/biochemie/trna/	tRNA viewer and sequence editor
UTRdb/UTRsite	http://bighost.area.ba.cnr.it/srs6/	5′-/3′-UTRs of eukaryotic mRNAs

Note: Abbreviations used: NC, non-coding; UTR, untranslated region.

Genome database mining refers to the process of computational genome annotation by which an uncharacterized DNA sequence is documented by the location along the DNA sequence involved in genome functionality. Two levels of computational genome annotation are structural annotation or gene identification and functional annotation or assignment of gene functionality. Structural annotation refers to the identification of ORFs and gene candidates in a DNA sequence using the computational gene discovery algorithm (Borodovsky and McIninch, 1993; Guigo *et al.*, 1992; Huang *et al.*, 1997). Functional annotation refers to the assignment of function to the predicted genes using sequence similarity searches against other genes of known function (Bork *et al.*, 1998; Gelfand, 1995).

There are two approaches in the development of sequence analysis techniques for gene identification (Fickett, 1996; Brent and Guigó, 2004): *ab initio* and *de novo* methods. The *ab initio* gene prediction is based only on the genome sequences using the statistical

TABLE 15.8 Some nucleic acid secondary databases

Database	URL	Description
Gene, gene motifs		
ACLAME	http://aclame.ulb.ac.be/	Genetic mobile elements
CORG	http://corg.molgen.mpg.de/	Conserved non-coding sequences
DEG	http://tubic.tju.edu.cn/deg	Essential genes from bacteria and yeast
EGO	http://www.tigr.org/tdb/tgi/ego/	Eukaryotic orthologous DNA sequences
Entrez Gene	http://ncbi.nlm.nih.gov/Entrez	Gene-centered information
Hoppsigen	http://pbil.univ-lyon1.fr/databases/hoppsigen.html	Human, mouse homologous processed pseudogenes
Imprinted gene cat.	http://www.otago.ac.nz/IGC	Imprinted genes in animals
MICdb	http://www.cdfd.org.in/micas	Prokaryotic microsatellites
STRBase	http://www.cstl.nist.gov/div83/strbase/	Short tandem DNA repeats
UniGene	http://www.ncbi.nlm.nih.gov/UniGene	Non-redundant eukaryotic gene-oriented clusters
Vector DB	http://geneom-www2.stanford.edu/vectordb/	Nucleic acid vectors
Xpro	http://origin.bic.nus.edu.sg/xpro/	Eukaryotic protein-encoding DNA
Exons, introns, splice sites		
ExInt	http://sege.ntu.edu.sg/wester/exint/	Exon-intron structures
HS3D	http://www.sci.unisannio.it/docenti/rampone/	Human splice sites
IDB/IEDB	http://nutmeg.bio.indiana.edu/intron/index.html	Intron and evolution
ISIS	http://isis.bit.uq.edu.au/front.html	Intron sequences
SpliceDB	http://www.softberry.com/berry.phtml?topic=	Mammalian splice sites
SpliceNest	http://splicenest.molgen.mpg.de/	View gene splicing from EST
Transcription/regulation (sites and elements/factors)		
ACTIVITY	http://wwwmgs.bionet.nsc.ru/mgs/systems/activity/	Functional D/RNA site activity
JASPER	http://jaspar.cgb.ki.se	TF DNA-binding sites
MAPPER	http://bio.chip.org/mapper	Putative TF binding sites
TESS	http://www.cbil.upenn.edu/tess	TE search system
TRANSFAC	http://transfac.gbf.de/TRANFAC/index.html	TF and binding sites
TRED	http://rulai.cshl.edu/tred	Transcription regulatory elements
DoOP	http://doop.abc.hu/	Orthologous promoters, plants
DPInteract	http://arep.med.harvard.edu/dpinteract	DNA-binding protein sites, *E. coli*
EPD	http://www.epd.isb-sib.ch	Eukaryotic promoters
PLACE	http://www.dna.affrc.go.jp/htdocs/PLACE	*cis*-Acting regulatory DRE, plants
PlantCARE	http://intra.psb.ugent.be:8080/PlantCARE/	Promoters & DRE, plants
PlantProm	http://mendel.cs.rhul.ac.uk/	Plant promoter seq for RNA pol II
PRODORIC	http://prodoric.tu-bs.de/	Prokaryotic gene regulation NW
DBTBS	http://dbtbs.hgc.jp/	Promoter & TF, *Bacillus subtilis*
PromEC	http://bioinfo.md.huji.ac.il/marg/promec	Promoters, *E. coli*
TRRD	http://www.bionet.nsc.ru/trrd/	Eukaryotic TE/F-binding regions

Note: Abbreviations used: cat, catalogue; NW, networks; TE/F, transcription elements/factors.

features of different regions/signals in genomic sequences. This method attempts to compose more or less concise descriptions of prototype objects and then identify genes by matching to such prototypes. The method identifies statistical patterns that distinguish coding from noncoding DNA sequences. A good example is the use of consensus sequences in identifying promoter elements or splice sites, such as DOUBLESCAN

TABLE 15.9 Some useful genomic analysis servers

Server	URL	Description
	Sequence alignment/analysis	
DIALIGN	http://bibiserv.techfak.uni-bielefeld.de/dialign/	Multiple sequence alignment
MAVID	http://baboon.math.berkeley.edu/mavid/	Multiple sequence alignment
MultiPipMaker	http://bio.cse.psu.edu	Multiple genomic seq align/analysis
ECR Browser	http://ecrbrower/dcode.org/	Comparative genomics
VISTA	http://www-gsd.lbl.gov/vista/	Comparative genomics
Theatre	http://www.hgmp.mrc.ac.uk/Registered/	Comparative genomic seq analysis
FAN	http://bioinf.man.ac.uk/cgi-bin/neil/ntfront.pl	Fingerprint analysis of nt sequences
	Gene, gene motifs	
CSTminer	http://www.caspur.it/CSTminer/	Identification of CDS/NCDS tags
GeneFizz	http://pbga.pasteur.fr/GeneFizz	Comparison of CDS/NCDS
e2g	http://bibiserv.techfak.uni-bielefeld.de/e2g/	Mappin EST/cDNA to genomic seq
ESTAnnotator	http://genome.dkfz-heidelberg.de	EST annotation
MGAlignIt	http://origin.bic.nus.edu.sg/mgalign/mgalignit	Genomic, mRNA/EST seq align
Cluster-Buster	http://zlab.bu.edu/cluster-buster/	DNA dense clusters
ACMES	http://acmes.rnet.missouri.edu/	Short repeat sequence search
SIC	http://stat.genopde.cnrs.fr/SIC/	Short inverted segments
Mreps	http://www.loria.fr/mreps/	DNA tandem repeats
TRACTS	http://bip-weizmann.ac.il/miwbin/servers/tracts	Oligo-Pu, -Py & other binary tracts
PatSearch	http://bighost.area.ba.cnr.it/BIG/PatSearch	Pattern/structural motifs
ERPIN	http://tagc.univ-mrs.fr/erpin/	Identification of RNAmotifs
Asmodeler	http://genome.ewha.ac.kr/Ecgene/ASmodeler	Modeling of alternative splicing
	SNP analysis	
CLOURE	http://imtech.res.in/~anand/cloure.html	SNP analysis by Clustal
PARSESNP	http://www.proweb.org/parsesnp	SNP analysis
PupaSNP	http://pupasnp.bioinfo.cnio.es/	SNP search
	Transcription, gene regulation	
PromoSer	http://biowulf.bu.edu/zlab/promoSer	Mamm promot/tscriptional start site
MATCH$^{\text{TM}}$	http://compel.bionet.nsc.ru/match/match.html	TF binding site search
Gibbs RS	http://bayesweb.wadsworth.org/gibbs/gibbs.html	TF binding site search
YMF	http://bio.cs.washington.edu/software	Discovery of TF binding site
POBO	http://ekhidna.biocenter.helsinki.fi9801/pobo	TF binding site verification
MSCAN	http://mscan.cgb.ki.se/cgi-bin/mscan	Clusters of TF identification
SiteSeer	http://rocky.bms.umist.ac.uk/SiteSeer/	Analysis of TF binding site
FIE2	http://sdmc.lit.org.sg/FIE2.0	Human gene translation start site
CARRIE	http://zlab.bu.edu/CARRIE-web	Regulatory network inference
	RNA Interference	
DEQOR	http://cluster-1.mpi-cbg.de/Deqor/deqor.html	Design of siRNAs
siDirect	http://design.rnai.jp/	SiRNA design for mammal RNAi
siRNA Selection	http://jura.wi.mit.edu/bioc/siRNA/	Prediction of siRNA oligo
TROD	http://www.cellbio.unige.ch/RNAi.html	T7 RNAi oligo design
	Structure, folding	
mfold	http://www.bioinfo.rpi.edu/applications/mfold	Nucleic acid folding
Sfold	http://sfold.wadsworth.org/	Nucleic acid folding & design
Pfold	http://www.daimi.au.dk/^compbio/pfold	RNA sec structure prediction
Rdfolder	http://ma.cbi.pku.edu.cn/	RNA sec structure prediction
RNAsoft	http://www.RNAsoft.ca	RNA sec structure prediction
ILM	http://cic.cs.wustl.edu/RNA/	RNA sec structure with pseudoknots
CARNAC	http://bioinfo.lifl.fr/carnac/	RNA folding family
GPRM	http://bioinfo.cis.nctu.tw/service/gprm	Common RNA sec structures
THERMODYN	http://wings.buffalo.edu/gsa/dna/dk/WEBTHERMODYN/	DNA helical stability profiling
MolProbity	http://kinemage.biochem.duke.edu	Structure validation

Note: Abbreviations used: align, alignment; sec, secondary; seq, sequence(s); promot, promoter; Pu, purine; Py, pyrimidine; oligo, oligonucleotide(s); TF, transcription factor; CDS/NCDS, coding sequence/non-coding sequence tags; nt, nucleotides; mamm, mammalian.

(Meyer and Durbin, 2002), AUGUSTUS (Stanke *et al.*, 2004). Curation of data is a prerequisite to developing pattern recognition algorithms for identifying features of biological interest for which well-cleansed, nonredundant databases are needed. Information on nonredundant, gene-oriented clusters can be found from UniGene of NCBI (http://www/ncbi.nlm.nih.gov/UniGene/) and TIGR Gene Indices at http://www.tigr.org/tdb/tdb.html (Lee *et al.*, 2005).

The *de novo* gene finding programs align homologous sequences and annotate them with respect to coding and conserved non-coding regions. Thus *de novo* gene predictions work by identifying gene pattern differences between the various regions, such as coding potential, splice signals and exon length distributions. This method attempts to identify a gene or gene component by finding a similar known object in available databases and relies on finding coding regions by their homology to known expressed sequences. An excellent example is the search for genes by trying to find a similarity between the query sequence and the contents of the sequence databases such as GeneID (Guigó *et al.*, 1992), Genie (Reese *et al.*, 1997), GenScan (Burge and Karlin, 1997), SLAM (Alexandersson *et al.*, 2003). Table 15.10 lists available on-line gene identification programs.

Some representative human gene expression databases include Cellular Response Database (http://LH15.umbc.edu/crd), GeneCards (http://bioinformatics.weizmann.ac.il/cards/), Globin Gene Server (http://globin.csc.psu.edu), and Human Developmental Anatomy (http://www.ana.ed.ac.uk/anatomy/database/). Although the accuracy for compact genome prediction is relatively high, the accuracy for mammalian genomes has lagged behind owing to the presence of a large number of pseudogenes and small fraction of coding sequences.

The basic information flow for an overall gene identification protocol follows:

15.3.1 Masking repetitive DNA

A map of repeat locations shows where regulatory and protein-coding regions are unlikely to occur (Smit, 1996). It is therefore best to locate and remove interspersed and simple

TABLE 15.10 Web-based gene identification programs

Program	URL of web site	Ref.
AGenDA	http://www.hgmp.mrc.ac.uk/GenomeWeb/nuc-geneid.html	1
AUGUSTUS	http://augustus.gobics.de	2
BLASTX	http://www.ncbi.nlm.nih.gov/BLAST/	3
EU*GENE*HOM	http://genopole.toulene.inra.fr/fioinfo/eugene/EuGeneHom/cgi-bin/EuGeneHom.pl	4
GeneFinder	http://dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html	5
GeneID	http://www1.imm.es/software/geneid/geneid.html	6
Genie	http://www.fruitfly.org/seq_tools/genie.html	7
GenScan	http://bioweb.pasteur.fr/seqanal/interfaces/genscan.html	8
Grail	http://compbio.ornl.gov/Grail-1.3/Toolbar.html	9
HMMgene	http://www.cbs.dtu.dk/services/HMMgene/	10
SGP-2	http://www1.imim.es/software/sgp2/	11
SLAM	http://bio.math.berkeley.edu/slam	12
TWINSCAN	http://genes.cs.wustl.edu/	13
WebGenMark	http://dixie.biology.gatech.edu/GeneMark/hmmchoice.html	14

Note: The computational techniques/web sites are quoted form the references: [1] Taher *et al.* (2004); [2] Stanke *et al.* (2004); [3] McGinnis and Madden (2004); [4] Foissac *et al.* (2003); [5] Solovyev *et al.* (1994); [6] Guigó *et al.* (1992); [7] Reese *et al.* (1997); [8] Burge and Karlin (1997); [9] Uberbacher *et al.* (1991); [10] Krogh (1977); [11] Parra *et al.* (2003); [12] Alexandersson *et al.* (2003); [13] Korf *et al.* (2001); [14] Lukashin and Borodovsky (1998).

repeats from eukaryotic sequences as the first step in any gene identification analysis. Because such repeats rarely overlap promoters or the coding portions of exons, their locations can provide important negative information on the location of gene features. RepeatMasker (http://www.genome.washington.edu/analysistools/repeatmask.htm) provides annotation and masking of both interspersed and simple repeats.

15.3.2 Database searches

Sequence similarity to other genes or gene products provides strong positive evidence for exons. Thus searching for a known homologue is perhaps the most widely understood means of identifying new CDS. The homologous sequence search can be conducted at one of INSDC sites. For protein-coding genes, translating the sequence in all six possible reading frames and using the result as a query against databases of amino acid sequences and functional motifs is usually the best first step for finding important matches (Gelfand *et al.*, 1996). It is shown (Green *et al.*, 1993) that:

(a) most ancient conserved regions (ACRs), which are regions of protein sequences showing highly significant homologies across phyla, are already known and may be found in current databases;

(b) roughly 20–50% of newly found genes contain an ACR that is represented in the databases; and

(c) rarely expressed genes are less likely to contain an ACR than moderately or highly expressed ones.

The EST databases probably contain fragments of a majority of all genes (Aaronson *et al.*, 1996). Current estimates are that the public EST databases of human and mouse ESTs represent more than 80% of the genes expressed in these organisms. Thus they are an important resource for locating some part of most genes. Finding good matches to ESTs is a strong suggestion that the region of interest is expressed and can give some indication as to the expression profiles of the located genes. However, the problem of inferring function remains because an EST that represents a protein without known relatives will not be informative about gene function. An access to dbEST can be found at http://www.ncbi.nlm.nih.gov/dbEST/index.html (Table 15.3).

15.3.3 Codon bias detection

Statistical regularity suggesting apparent 'codon bias' over a region is one of the indicators of protein-coding regions. Information on codon usage can be found at the Codon usage database, http://www.kazusa.or.jp/codon/ or http://biochem.otago.ac.nz:800/Transterm/homepage.html. Most computational identification of protein-coding genes relies heavily on recognizing the somewhat diffuse regularities in protein-coding regions that are due to bias in codon usage. Some of the most informative coding measures are:

- dicodon counts i.e. frequency counts for the occurrence of successive codon pairs;
- certain measure of periodicity, e.g. the tendency of multiple occurrences of the same nucleotide to be found at distances of 3n bp;
- a measure of homogeneity versus complexity, i.e. counting long homopolymer runs; and
- open reading frame occurrence (Fickett and Tung, 1992).

Many coding region detection programs are primarily the result of combining the numbers from one or more coding measures to form a single number, called a discriminant. Typically the discriminant is calculated in a sliding window (i.e., for successive subsequence of fixed length) and the result is plotted. To gain significant information from a coding measure discriminant, a DNA chain of more than 100 bases is required.

15.3.4 Detecting functional sites in the DNA

Because matches to template patterns may indicate the location of functional sites on the DNA, it would be more assuring if we were able to recognize the locations such as transcription factor binding sites and exon/intron junctions where the gene expression machinery interacts with the nucleic acids. One way to summarize the essential information content of these locations, termed signals. is to give the consensus sequence, consisting of the most common base at each position of an alignment of specific binding sites. Consensus sequences are useful as mnemonic devices but are typically not reliable for discriminating true sites from pseudosites. One technique for discriminating between true sites and pseudosites with a basis in physical chemistry is the position weight matrix (PWM). A score is assigned to each possible nucleotide at each possible position of the signal. For any particular sequence, considered as a possible occurrence of the signal, the appropriate scores are summed to give a score to a potential site. Under some circumstances this score may be approximately proportional to the energy of binding for a control ribonucleoprotein/protein. The following functional sites/signals are considered:

Promoters. The promoter is an information-rich signal that regulates transcription especially for eukaryotes. Computer recognition of promoters (Fickett and Hatzigeorgiou, 1997) is important partly for the advance it may provide in gene identification. Available programs include those depending primarily on simple oligonucleotide frequency counts and those depending on libraries describing transcription factor binding specificity's, together with some description of promoter structure. Information on eukaryotic polymerase II (mRNA synthesis) promoters can be found from EPD (Praz *et al.*, 2002) at http://www.epd.isb-sib.ch and PlantProm (http://mendel.cs.rhul.ac.uk/) for plants (Table 15.8).

Intron splice sites. The position weight matrices have been complied for splice sites in a number of different taxonomic groups (Senapathy *et al.*, 1990) and these may be the best resource available for analysis in many organisms. Unfortunately, PWM analysis of the splice junction provides rather low specificity perhaps because of the existence of multiple splicing mechanisms and regulated alternative splicing. Many of the integrated gene identification services provide separate splice site predictions. ExInt database (http://intron.bic.nus.edu.sg/exint/extint.html) provides information on the exon-intron structure of eukaryotic genes. Various information resources related to introns can be found at EID (http://mcb.harvard.edu/gilbert/EID/) and Intron (http://nutmeg.bio.indiana.edu/intron/index.html) servers.

Translation initiation site. In eukaryotes, if the transcription start site is known, and there is no intron interrupting the 5′ UTR, Kozak's rule (Kozak, 1996) probably will locate the correct initiation codon in most cases. Splicing is normally absent in prokaryotes, yet because of the existence of multicitronic operons, promoter location is not the key information. Rather the key is reliable localization of the ribosome binding site. The TATA sequence about 30 bp from the transcription start site may be used as a possible resource.

Termination signals. The polyadenylation and translation termination signals also help to demarcate the extent of a gene (Fickett and Hatzigeorgiou, 1997).

Untranslated regions. The 5′ and 3′ UTRs are portions of the DNA/RNA sequences flanking the final coding sequence. They are highly specific both to the gene and to the species from which the sequences are derived. Thus UTRs can be used to locate CDS, e.g., eukaryotic polyadenylation signal (Wahle and Keller, 1996), mammalian pyrimidine-rich element (Baekelandt *et al.*, 1997), mammalian AU-rich region (Senterre-Lesenfants *et al.*, 1995), *E. coli* fbi (3′-CCGCGAAGTC-5′) element (French *et al.*, 1997) etc. Information on 5′ and 3′ UTRs of eukaryotic mRNAs can be found at UTRdb (http://bigrea.area.ba.cnr.it:8000/srs6/) (Pesole *et al.*, 2002).

Computational analyses of DNA sequences/gene identification can be carried out by the use of Internet resources. Online bibliographies (indexed by years) on computational gene recognition are maintained at http://linkage.rockefeller.edu/wli/gene/right.html. Commonly used computational gene discovery programs accessible through the web are tabulated in Table 15.10. Most programs are organism specific and require a selection of organisms. Since no single tool can perform all the relevant sequence analysis leading to gene identification, it is recommended to submit the query sequence to the analysis of several software packages (tools) to make use of the best computational techniques. It must be emphasized that a gene predicted by computational methods must be viewed as a hypothesis subject to experimental verification.

15.4 GENE EXPRESSION

15.4.1 Expression profiling: DNA chips [Adapted from Schena 2003]

Advances in microarray technology (Müller and Nicolau, 2005; Schena, 2003) have increased the speed at which sequence information and differential gene express data can be gathered (Stekel, 2003; Young, 2000). Organisms express their genes at a relatively constant rate until the products encoded by those genes are needed for specific function. When needed, genes are activated or repressed rapidly and in dramatic fashion changing by hundred- or thousand-fold, depending on the particular gene and the strength of the regulatory cue. The expression of genes changes in response to a wide spectrum of signals providing a gene expression ‘fingerprint’ that is characteristic for a given physiological state. Because gene expression correlates specifically and tightly with function, it is possible to infer the function of genes and the interaction of pathways by documenting which gene are turned up or down in a given physiological state. Gene regulation provides a selective evolutionary advantage by conserving cellular building blocks and enzymatic machinery until they are needed, and by allowing the organism to adapt to a plethora of different environmental conditions to which it is exposed during its lifetime. Disease states, drug treatment, different developmental stages and many other processes can be examined by cataloging gene expression profiles.

The DNA microarrays (also known as DNA chips or gene chips) allow the analysis of the entire genome from an organism in a single experiment (Bowtell, 1999). DNA chips made by the delivery approach deliver target density of 100–10 000 elements/cm^2 and oligonucleotides of up to 120 bases, while the synthetic method achieves a higher target density in the range of 100–500 000 elements/cm^2 with oligonucleotides of up to 25 bases.

The widespread use of PCR amplification in microarray manufacture places much importance on the availability of informatics tools for automated primer design. A primer is a synthetic oligonucleotide that hybridizes to a complementary nucleic acid template, and expedites (primes) enzymatic synthesis by providing a starting point for enzyme. PCR primers bind to proximal and distal sites on cDNA templates, allowing target sequences to be amplified with DNA polymerase in the PCR process. There are a number of design considerations when choosing PCR primers, including length, melting temperature, sequence composition, secondary structure and position with the DNA template.

The application of photolithography to oligonucleotide synthesis for DNA chips (Fodor *et al.*, 1991), as illustrated in Figure 15.4, begins with a glass substrate modified with silane reagents to provide a surface containing reactive amine groups. The reactive amino groups are modified with a second reagent that contains a specific chemical group, mehylnitropiperonyloxycarbonyl (MeNPOC). The MeNPOC group is stable to variety of chemical reagents but can be removed selectively by illuminating UV on the chip for a period of about 30 s. MeNPOC group prevents chemical reactions from taking place in the absence of UV light and are hence known as photoprotecting groups. When the photoprotecting groups have been removed, the deprotected regions on the surface can react efficiently with a special family of DNA nucleotides to form chemical bonds at the 3′ position of the deoxyribose, which contains a reactive phosphoramidite group. Each base that has undergone coupling has a MeNPOC photoprotecting group on its 5′ hydroxyl position. The MeNPOC group on the coupled base can be removed by exposure to UV light, allowing coupling to a second base. A repeated series of deprotection and coupling steps allows oligonucleotides of any sequence to be synthesized on the glass surface in a stepwise manner.

Oligonucleotide synthesis can be directed to discrete location on the microarray surface using photomasks, which allow the selective deprotection of certain regions of the chip. The photomasks contain chrome-plated glass with interspersed clear region. Chrome prevents the passage of UV, while the clear regions allow light to pass and impinge on the chip surface. Because the masks can be manufactured to contain any pattern of chrome and clear regions, it is possible to direct light activation to any region of the chip in any desired order. A series of photomasks can thus be used to direct the stepwise synthesis of

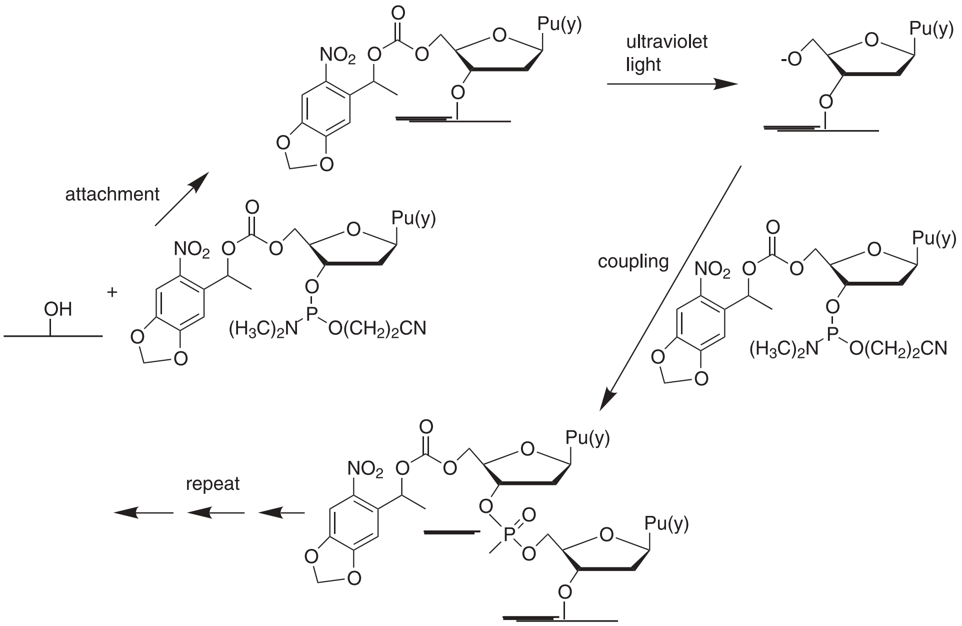


Figure 15.4 Photolithography of oligonucleotide synthesis. Adapted from Schena (2003)

oligonucleotide microarrays containing any sequence of interest, with each mask directing the synthesis of one DNA base at each location on the substrate. The chrome checkerboards can be highly miniaturized; producing microarray features sizes in the 20–50 µm range. The 1.28 cm^2 chip is enclosed in a plastic cassette that prevents it from being damaged and provides an extremely convenient hybridization chamber. Microarrays made by photolithography and other semiconductor-based strategies typically produce 15–30 µm features, and printed microarray spot size is generally 50–350 µm. A typical printed DNA spot contains approximately 1 billion (10^9) molecules attached to the glass substrate. The DNA or synthetic oligonucleotide sequences are arrayed on glass chips. By labeling probe cDNA/mRNA with different fluorescent labels (for different states), the RNA expression profiles of different states may be rapidly performed through simultaneous hybridization to an appropriate microarray (DeRisi *et al.*, 1996; Perou *et al.*, 1999; Wang *et al.*, 1999).

The target-probe hybridization interactions are most prominently used in nucleic acid microarray analysis. Many factors (Table 15.11) affect hybridization efficiency and the strength of the duplex. Hybridization affinity and signal intensity correlates directly in microarray assays, with the G:C sequences producing extremely intense fluorescence, the A:T(U) sequences producing much weaker fluorescence. Sequence composition is a minor consideration if the targets and probes are long. In most gene expression studies, homology is 100% because target sequences hybridize to probe molecules derived from their cognate genes. Cross-hybridization occurs if gene families that are highly conserved or have a number of members hybridizing to noncognate targets and can complicate gene expression measurement. Cross-hybridization is observed if the sequence identity between target and probe is ≥70% but is generally not a factor for sequences share <70% nucleotide identity.

TABLE 15.11 Factors affecting hybridization in DNA microarray analysis

Parameter	Effects	Remarks
Base pair	G:C pair has stronger hybridization affinity than A:T(U) pair	The difference between G:C and A:T(U) can be minimized by the use of buffer containing tetramethylammonium chloride
Homology	The higher the homology the better hybridization, i.e. higher efficiency	Week or no signal is produced under stringent conditions for less than 70% homology
Temperature	Elevated temperature improves hybridization efficiency by speeding diffusion and reducing secondary structure, however excessive temperature reduces the efficiency by melting duplex	Optimal hybridization temperature is ~10°C below the melting temperature of the heteroduplex.
Salt	Salt improves hybridization efficiency by minimizing electrostatic repulsion	Most hybridization buffers use [Na$^+$] in the range of 0.4–1.0 M
pH	Neutral pH favors hybridization by promoting hydrogen bond formation between bases	Optimal pH is in the range of 5.5–8.5.
Size of target molecule	Reducing the size can improve hybridization specificity by reducing or eliminating cross-hybridization	Reducing size can be important in gene expression experiments involving highly conserved gene families
Size of probe molecule	Reducing size can improve hybridization efficiency by speeding diffusion and reducing secondary structure formation	Size can be reduced by mechanical shearing, enzymatic digestion or chemical cleavage

Taken from Schena (2003)

Long heteroduplexes have a much greater affinity than short heteroduplexes. For gene bps are commonly used. For genotyping applications, single base discrimination is the key and so shorter target-probe interactions (e.g. 12–20 bp) are often employed. The extent of complementarity between target and probe is an important determinant of hybridization affinity. For a perfect heteroduplex (100% complementarity) sharing 15 bp, an average hybridization temperature in a standard hybridization buffer is 42°C, which is approximately 10°C below the melting temperature. For each additional bp added, the hybridization temperature must be elevated by 1–2°C to maintain sufficient hybridization. Hybridization involving 25mer heteroduplexes typically requires hybridization temperature of ~60°C.

The capacity to distinguish single-base differences between targets and probes forms the basis of genotyping applications of microarrays. Most human diseases result from a single-base change in a critical region of the coding sequence, which leads to the synthesis of a defective protein. Most patients afflicted with a genetic condition can be distinguished from the normal population via probes that hybridize with high affinity to the wild type locus and with lesser affinity to the disease locus. To maximize the capacity to distinguish single-base differences, the heteroduplexes must be relatively short (e.g. 15 bp) and the mutation located in the center position of the heteroduplex. Reagents that destabilize base pairing are used in some microarray hybridization reactions. Formamide reduces the melting temperature of hybridized sequences. Addition of 50% formamide reduces the melting temperature by ~25°C and improves the detection by reducing background fluorescence. Tetramethylammonium chloride selectively reduces the strength of G:C bonds relative to A:T(U) bonds and thus minimizes binding differences between G:C-rich and A:T-rich sequences. This property is useful for genotyping applications in which a similar signal for each heteroduplex, irrespective of the sequence composition, improves the quality of the assay.

The complete genomic sequence of a eukaryotic organism provides a linear map of the genes along each chromosome. If the position of each gene is known on the microarray and in the genome, microarray data can be superimposed on the genome to provide an expression map. An expression map is the graphical representation in which gene expression values are superimposed on to their cognate genes along the chromosomes. It allows the researcher to visualize transcript abundance and regulation with respect to their physical location of the chromosomes. Expression data can also be superimposed on to biochemical and genetic pathways to correlated gene expression with metabolic activity (DeRisi *et al.*, 1996), known as pathway analysis.

The amount of genomic information on a microarray is expressed as microarray complexity, calculated as the product of the number of unique sequences, times their average length in nucleotide, i.e.:

$$\text{Microarray complexity} = \text{Number of unique sequences} \times \text{Average length}$$

The complexity of a microarray is often expressed as a function of the size of genome represented in terms of a genome equivalent, i.e.:

$$\text{Genome equivalent} = \text{Microarray complexity}/\text{Genome size}$$

The standard printing convention uptakes samples from the upper left corner to the right lower corner of the microplate designated as A1 to Xn (where letter and number represent row and column respectively). The electronic file that contains all of the information pertinent to each microarray spot is known as a content map. Content maps include microplate location, microarray number, sequence information, GenBank accession number, gene names and other important data, as well as normalized intensity values obtained through quantitation.

One of the aims of microarray analysis in gene expression profiling is to build comprehensive gene expression databases for each organism that contain expression profiles for each gene across thousands of different conditions, thereby allowing biological exploration to take place predominantly by means of a computer. Various microarray databases and servers (Table 15.12) are available in support of expression profiling.

TABLE 15.12 Microarray databases and servers

Database/Server	URL	Description
Databases		
ArrayExpress	http://www.ebi.ac.uk/arrayexpress	Collection of MA GE data
BodyMap	http://bodymap.ims.u-tokyo.ac.jp/	Human and mouse GE data
BGED	http://love2.aist-nara.ac.jp/BGED	Brain GE data
CleanEx	http://www.cleanex.isb-sib.ch/	Expression reference, data compar
DbERGEII	http://dberge.cse.psu.edu/menu.html	GE, genomic align, annotation
EICO DB	http://fantom2.gsc.riken.jp/EICODB/	Discovery of novel imprinted genes
EPConDB	http://www.cbil.upenn.edu/EPConDB	Endocrine pancreas consortium
GeneNote	http://genecards.weizmann.ac.il/genenote/	Human GE in healthy tissues
GenePaint	http://www.genepaint.org/Frameset.html	Mouse GE patterns
GeneTide	http://genecards.weizmann.ac.il/genetide/	GeneCards transciptome-members
GEO	http://www.ncbi.nlm.nih.gov/geo/	GE omnibus: GE profiles
GermOnline	http://www.germonline.org/	GE in mitotic and meiotic cell cycle
GXD	http://www.informatics.jax.org/	Mouse GE
H-ANGEL	http://jbirc.aist.go.jp/hinv/index.jsp	Human anatomic GE library
HemBase	http://hembase.niddk.nih.gov/	GE in diff human erythroid cells
HugeIndex	http://hugeindex.org/	Human GE levels in normal tissues
LOLA	http://www.lola.gwu.edu/	Compare gene sets from different expts
MethDB	http://www.methdb.de/	DNA methyl data, patterns, profiles
Oste-Promoter DB	http://www.opd.tau.ac.il	Genes in osteogentid prolif and diff
PEDB	http://www.pedb.org/	Prostate GE:EST from specif cDNA
PEPR	http://microarray.cnmcresearch.org/pgadatable.asp	Public GE resource: profiles in a variety of diseases and conditions
RECODE	http://recode.genetics.utah.edu/	GE by programmed translation
RefExA	http://www.lsbm.org/db/index_e.html	Reference for human GE analysis
Stanford MA DB	http://genome-www.stanford.edu/microarray	Raw and normalized MA data
yMGV	http://www.transcriptome.ens.fr/ymgv/	Yeast MA global viewer
Servers		
GEPAS	http://bioinfo.cnio.es	MA gene expression data analysis
GenePublisher	http://www.cbs.dtu.dk/services/GenePublisher	DNA MA data analysis
ExpressYourself	http://bioinfo.mbb.yale.edu/expressyourself	Processing/visualization MA data
ChipInfo	http://biosun1.harvard.edu/complab/chipinfo/	Gene annotation for MA analysis
REDUCE	http://bussemaker.bio.columbia.edu/reduce/	Transcriptional activities from MA
KARMA	http://ymd.med.yale.edu/karma/cgi-bin/karma.pl	Heterogeneous MA platform compar
ArrayProspector	http://string.embl-heidelberg.de8080/	MA expression data inference
ArrayPipe	http://www.pathogenomics.ca/arraypipe/	MA data processing
ArrayXPath	http://www.snubi.org/software/ArrayXPath	Mapping MA gene expression data
Expression Profiler	http://ebi.ac.uk/expressionprofiler	MA data analysis
ProbeLynx	http://www.pathogenomics.ca/probelynx	MA gene probe
CONFAC	http://morenolab.whitehead.emory.edu/cgi-bin/confac/login.pl	Genomic promoter analysis and MA

Note: Abbreviations used: compar, comparison; DB, database; GE, gene expression; MA, microarray; diff, differentiating; prolif, proliferation; methyl, methylation.

15.4.2 Gene expression: mRNA quantification and transcriptome analysis

Nearly all the biological events are associated with changes in expression of key genes. By comparing gene expression profiles under different conditions, individual genes or group of genes can be identified that play an important role in a particular physiological cascade or process, all of which involve gene expressed products, RNA and proteins. It has been long recognized that mRNA plays a pivotal role in determining the nature and quantity of proteins produced by cells. The differences in quality and quantity of proteins of different cell types are a reflection of differences in the mRNA species expressed (a typical mammalian cell expresses ~15 000 different species of mRNA) and of their levels of expression during cellular development and maintenance. Technological advances in nucleic acid biochemistry contribute to the development of a range of techniques for quantifying mRNA on a genomic scale, sometimes referred to as transcriptome (Velculescu *et al.*, 1997). These include sequencing of cDNA libraries, differential display PCR (DD-PCR), serial analysis of gene expression (SAGE) and DNA array hybridization. The DNA array hybridization (DNA microarray), which has been described in the preceding subsection, uses cDNA or oligonucleotides as target molecules to assess the types and levels of labeled mRNA. The tool is applicable to genome-wide expression profiling experiments (Wodicka *et al.*, 1997; Richmond *et al.*, 1999). The array expression database can be accessed at EBI (http://www.ebi.ac.uk/arrayexpress/).

15.4.2.1 Sequencing of cDNA library. The preparation of a cDNA library has been described previously (subsection 15.1.4). The library commonly contains more than 1 million clones. Therefore by sequencing clones from a cDNA library and then counting the frequency of appearance of a particular gene sequence, it is possible to gain an estimate of the relative abundance of each mRNA transcript. Thus the accurate quantification of abundance of a transcript is a direct function of the number of cloned sequences produced from the cDNA library (Vasmatzis *et al.*, 1998). The NCBI provides web-based tools, xProfiler (http://www.ncbi.nlm.nih.gov/ncicgap/cgapxpsetup.cgi) and Digital Differential Display (http://www.ncbi.nlm.nih.gov/UniGene/ddd.cgi?ORG=Hs), which allow the comparison of sequence data derived from cDNA libraries in their collection. This approach can only apply to standard cDNA library information and not to normalized cDNA data.

15.4.2.2 Differential display PCR. Differential display PCR (DD-PCR) is a method for identifying cDNA fragments that are differentially expressed between two biological samples. It is based on generating cDNA fragments from mRNA using two oligonucleotide primers, one being complementary to the polyA tail of transcripts (e.g. oligo-dT_{11}) and the other a short random nucleotide sequence (e.g. dodecamer, dN_{10}). DD-PCR has the potential to identify all transcripts present in a biological sample when sufficient primer combinations are applied. After cDNA synthesis, the fragments are labeled during PCR amplification. The products are separated by hrPAGE on a sequencing gel and pattern of amplified cDNA fragments visualized. The intensity of the labeled band reflects the relative abundance of its mRNA transcript within the original mRNA population. Major differences in the cDNA band patterns generated from two biological samples using the same set of primers, indicate the presence of differentially expressed transcripts. Cloning and sequencing of the eluted cDNA bands enables the identity of the genes from which these cDNAs originate to be defined.

An improvement in DD-PCR can be made by the use of restriction enzyme digestion of double-stranded cDNA followed by ligation of an adapter to mediate selective PCR of the extreme 3′ ends of the fragmented cDNAs (Prasher and Weissman, 1996). This adaptation enables stringent PCR to be performed by specific primer annealing at 56°C (instead of 40°C for nonstringent primer), thus providing near-quantitative information on gene expression (Shiue, 1997; Renner *et al.*, 1998). The on-line primer design tool is accessible from CODEHOP (http://bioinformatics.weizmann.ac.il/blocks/codehop.html) and PDA (http://dbb.nhri.org.tw/primer/).

15.4.2.3 *Serial analysis of gene expression.* Serial analysis of gene expression (SAGE) is a high-efficiency method to simultaneously detect and measure the expression levels of genes expressed in a cell at a given time without a prior availability of transcript information (Pleasance *et al.*, 2003). The SAGE technique produces short unique 9–14 sequences (tags) that identify one (or more) mRNAs. The unique sequence tags are concatenated serially into long DNA molecules for full sequencing (Figure 15.5). The frequency of each tag in the concatenated sequence reflects the cellular abundance of the corresponding transcripts.

This serial analysis of many thousands of gene-specific tags allows the simultaneous accumulation of information for genes expressed in the tissue of interest and gives rise to an expression profile for tissue. Since SAGE uses a short sequence tag to distinguish each transcript, high-quality sequence data is essential for its accurate identification. Furthermore, the ability to successfully identify the originating transcript for a tag is directly related to the number of sequences deposited in databases for each species. An improvement involves the generation of 21 bp tags known as LongSAGE (Saha *et al.*, 2002). The comprehensive analysis of SAGE data requires software that integrates statistical data analysis with a database system. SAGE databases are accessible at SAGEmap (http://www.ncbi.nlm.nih.gov/SAGE) and SAGEnet (http://www.sagenet.org). To enhance the utility of SAGE data, SAGEGenie (http://cgap.nci.nih.gov/SAGE) provides an automatic link between gene names and SAGE transcript levels, analysis and dissemination of the gene expression profiles (Boon *et al.*, 2002).

15.4.2.4 *Systematic evolution of ligands by exponential enrichment.* Systematic evolution of ligands by exponential enrichment (SELEX) is a technique for isolating functional nucleic acids by screening large libraries of oligonucleotides via an iterative process of *in vitro* selection and amplification (Klung and Famulock, 1994). Single stranded nucleic acids fold into a bewildering number of stable conformations, available in the vast starting SELEX libraries. SELEX is normally use for the selective enrichment and amplification of the oligonucleotides (aptamers) with desired ligand binding characteristics. These oligonucleotides act as binding species (e.g. aptamers used as target molecules in nucleic acid microarrays) and catalytic nucleic acids (deoxyribozymes, ribozymes and aptazymes). However, these oligonucleotide aptamers work in a different manner compared to characteristic nucleic acid hybridization in that the ligand binding capacity is the result of the oligonucleotide's three-dimensional conformation, not nucleotide base–base complementarity. Thus these aptamers will bind to proteins and other molecules that do not normally interact with DNA or RNA (Jayasena, 1999).

The DNA/RNA library is constructed via combinatorial synthesis of oligonucleotides (containing constant sequences at the 5′- and 3′-ends and middle random sequences of generally 35–60 nucleotides). The complexity of the library derived from the combinatorial synthesis can be estimated. For example, in a library consisting of oligonucleotides of N nucleotides in length derived from 4 different nucleotides, the complexity (number of oligonucleotides) = 4^N. Normally the starting round for a SELEX process contains 10^{13}–10^{15} individual sequences. Essentially SELEX involves the repeated binding;

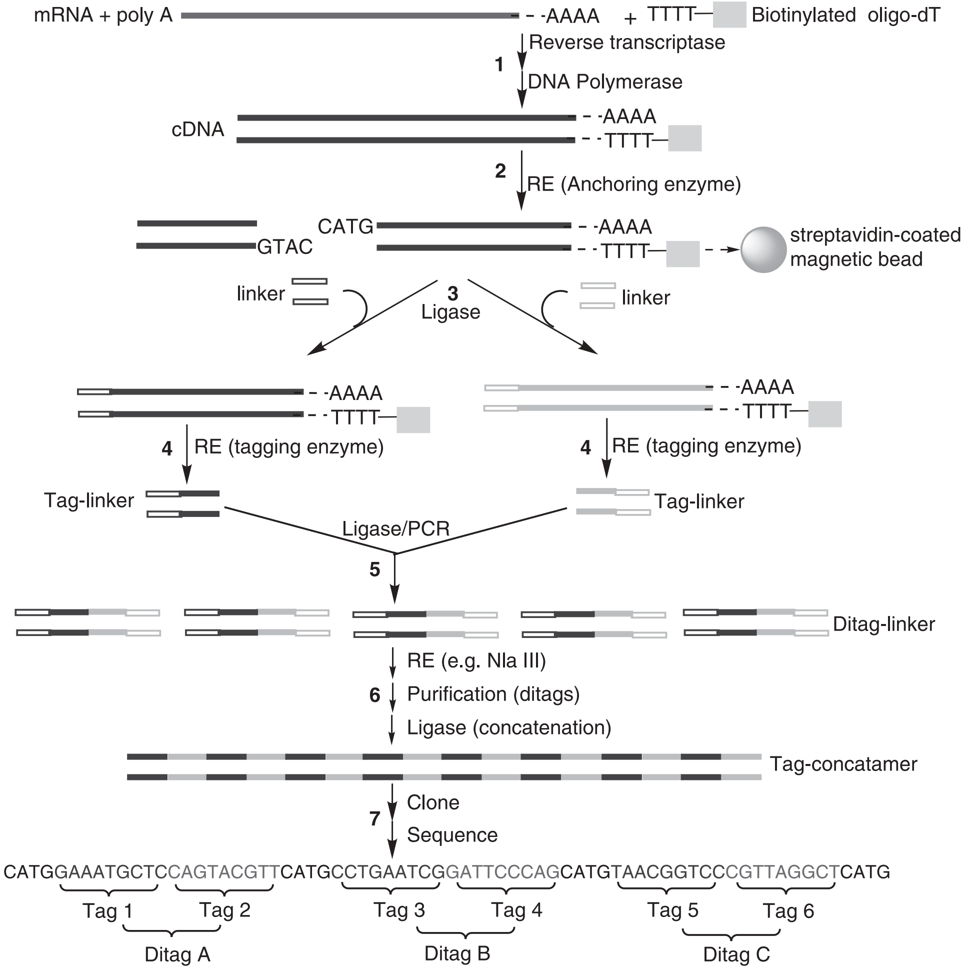

mRNA + poly A
- -AAAA
+ TTTT—
Biotinylated oligo-dT
1
Reverse transcriptase
DNA Polymerase
cDNA
- -AAAA
- -TTTT—
2
RE (Anchoring enzyme)
CATG
GTAC
- -AAAA
- -TTTT—
streptavidin-coated magnetic bead
linker
3
Ligase
linker
- -AAAA
- -TTTT—
- -AAAA
- -TTTT—
4
RE (tagging enzyme)
4
RE (tagging enzyme)
Tag-linker
Tag-linker
Ligase/PCR
5
Ditag-linker
RE (e.g. Nla III)
6
Purification (ditags)
Ligase (concatenation)
Tag-concatamer
7
Clone
Sequence
CATGGAAATGCTCCAGTACGTTCATGCCTGAATCGGATTCCCAGCATGTAACGGTCCCGTTAGGCTCATG
Tag 1
Tag 2
Tag 3
Tag 4
Tag 5
Tag 6
Ditag A
Ditag B
Ditag C

Figure 15.5 Scheme of SAGE method. Step 1: poly A + RNA is converted to cDNA using a biotinylated-oligoT primer. Step 2: the resulting cDNA is cleaved with a frequently cutting restriction enzyme known as anchoring enzyme (e.g. Nla III with 4 bp recognition site and cutting on average every 256 nt). Step 3: the 3′ end of the cDNA us captured by the use of streptoavidin-coated magnetic beads. The cDNA pool is then split in two (A and B) and separate linkers are ligated to each of the cDNA pools. The linkers contain a site for a type IIs restriction enzyme known as tagging enzyme because it cuts the cDNA at a site specific number of nucleotides away from the recognition site (e.g. Fok I cuts 13 bp downstream of its recognition site). Step 4: tagging enzyme cuts the cDNA/linker hybrid releasing a short sequence tag of cDNA attached to the linker. Step 5: the tag-linker hybrid molecules are ligated tail-to-tail and amplified by PCR *via* linker sequences. Step 6: linkers are removed from the ditags by the use of the anchoring enzyme. The ditags are purified and ligated together to form concatemers. Step 7: Concatemers are then cloned and sequenced to reveal the identity of each tag

selection and amplification of aptamers from the initial library until one or more displaying the desired characteristic have been isolated. The basic form of SELEX protocol (Figure 15.6) consists of the following series of steps:

1. A candidate mixture of nucleic acids of differing sequences is prepared. The candidate molecules consist of 5′- and 3′-regions of fixed sequences and the middle regions of randomized sequences. The fixed sequences are chosen either:

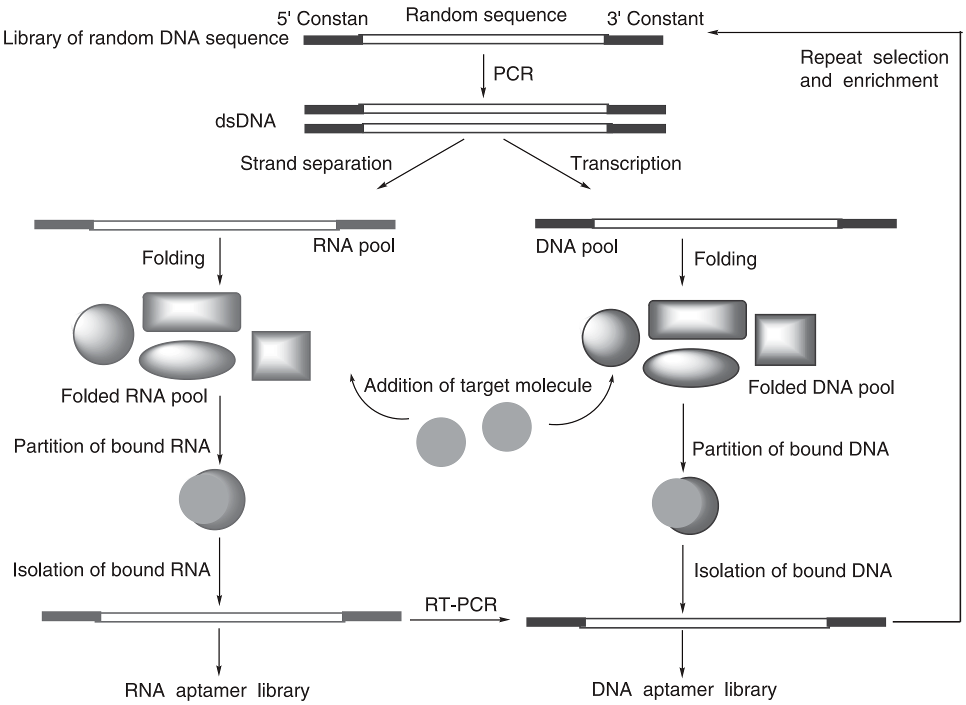

Library of random DNA sequence
5' Constan
Random sequence
3' Constant
Repeat selection and enrichment
PCR
dsDNA
Strand separation
Transcription
RNA pool
DNA pool
Folding
Folding
Folded RNA pool
Addition of target molecule
Folded DNA pool
Partition of bound RNA
Partition of bound DNA
Isolation of bound RNA
Isolation of bound DNA
RT-PCR
RNA aptamer library
DNA aptamer library

Figure 15.6 Scheme for systematic evolution of ligands by exponential enrichment. A library of DNA oligonucleotides containing randomized sequences is converted into dsDNA by PCR using constant 5′ and 3′ sequences for annealing primers. dsDNA is converted into ssDNA by strand separation or into RNA by *in vitro* transcription. Single stranded nucleic acids are allowed to fold into pools (DNA/RNA pool) of folded structures. The target molecule is mixed with nucleic acid pools and the non-specific or low affinity binding nucleic acid molecules are removed by washing. The captured DNA/RNA are eluted, recovered and amplified by PCR/RT (reverse transcriptase)-PCR to obtained a new enriched DNA library. The cycle is repeated until a specific population of DNA/RNA is finally obtained. For more than one molecular species of DNA/RNA, the DNA/RNA aptamer library can be constructed

a) to assist in the amplification steps;
b) to mimic a sequence known to bind the target; or
c) to enhance the concentration of a given structural arrangement of the nucleic acids in the mixture.

2. The candidate mixture is contacted with a selected target under the condition favorable for binding between the target and the members of the candidate mixture.
3. The nucleic acids with the highest affinity for the target are partitioned from those nucleic acids with lesser (or without) affinity to the target. It is generally desirable to set the partitioning criteria so that a significant amount of nucleic acids in the mixture (approximately 5–50%) are retained during partition.
4. Those nucleic acids selected during partitioning as having the high affinity for the target are then amplified to create a new candidate mixture.
5. The partition (selection) and amplification steps are repeated so that the candidate mixture contains fewer and fewer unique sequences with increased average degree of affinity for the retained nucleic acids.

Although SELEX was originally developed to screen oligonucleotide libraries for aptamers, the protocol is applicable for the identification of high affinity DNA/RNA sequences (Roulet *et al.*, 2002) to a wide variety of different targets including proteins, glycans and other biomolecules. The approach effectively extends DNA/RNA functionality beyond informational (complementarity) into conformational (binding and catalysis) realism. The oligonucleotide library can be made for SELEX by fragmentation of the DNA from an organism. In this application of genomic SELEX, regulatory proteins can be used to find those sequences that are the most likely *in vivo* targets of the proteins. Thus genomic SELEX offers the opportunity to identify the protein-oligonucleotide linkage map for an organism. SELEX_DB is accessible at http://wwwmgs.bionet.nsc.ru/mgs/systems/selex/ (Ponomarenko *et al.*, 2000) and aptamer database at http://aptamer.icmb.utexas.edu/ (Lee *et al.*, 2004).

15.5 GENOME PROJECT

The Genome On-Line Database (GOLD) at http://www.genomosonline.org is the resource site for information regarding complete or ongoing global genome sequencing projects (Liolios *et al.*, 2006).

The Human Genome Project (HGP) is a worldwide research initiative with the goal of analyzing the complete sequence of human DNA to identify all of the genes. A comprehensive description of the human genome is thus the foundation of human biology and the essential prerequisite for an in-depth understanding of disease mechanisms. With the introduction of somatic cell technology, recombinant DNA and polymorphic DNA markers, it becomes possible to dissect the human genome at the molecular level. The international HGP was officially launched in early 1990. Its general strategy can be described in three phases:

1. the generation of chromosome maps;
2. large-scale DNA sequencing; and
3. annotating the DNA sequence.

The human genome consisting 2.91 billion bp of DNA has been completely sequenced (Venter *et al.*, 2001)

Given the sequence information it will be possible to investigate the roles of all of the gene products, how they are controlled and interact, and their possible involvement in health and disease. The annotation of human DNA sequence will take several forms, including:

- cataloguing all of the genes;
- identifying the genes and DNA sequence variations that either directly cause or are associated with disease;
- studying genetic variation; and
- establishing integrated and curated databases containing all DNA sequence annotation.

At the present there are two general approaches to gene cataloguing (finding). The first involves detecting, isolating and analyzing transcripts of genes from genomic DNA. This can be accomplished by studying cDNAs, which are cloned copies of mRNA. An alternative approach involves computer-based analysis of the DNA sequence to search for structural features of genes such as start sites for gene transcription, protein coding regions and transcription stop sites (Fickett, 1996; Claverie, 1997). Part of this strategy can include comparing the DNA sequence of two organisms to search for evolutionarily conserved sequences to identify gene coding regions. A combination of both approaches to the

TABLE 15.13 Some genome project databases

Database	URL	Description
Human sequences, mapping and annotation		
Ensembl	http://www.ensembl.org/	Annotated euk genome information
AluGene	http://alugene.tau.ac.il/	Complete human Alu map
AllGenes	http://allgenes.org/	Human & mouse gene, transcript, protein
GDB	http://www.gdb.org/	Human gene/genomic maps
GenAtlas	http://www.genatlas.org/	Human genes, markers, phenotypes
GeneCards	http://bioinfo.weizmann.ac.il/cards/	Human genes, maps, proteins, diseases
GeneLoc	http://genecards.weizmann.ac.il/geneloc/	Gene locations
HOWDY	http://www-alis.tokyo.jst.jp/HOWDY/	Human organized whole genomes
HuGeMap	http://www.infobiogen.fr/services/Hugemap	Human genome genetic/physical map
IXDB	http://ixdb.mpimg-berlin-dahlem.mpg.de/	Physical map of human chrom X
Map Viewer	http://www.ncbi.nlm.nih.gov/mapview/	Genomic view by chrom position
MGC	http://mgc.nci.nih.gov/	Mammalian genome collection of ORF
mtDB	http://www.genpat.uu.se/mtDB	Human mitochondrial genomes
ParaDB	http://abi.marseille.inserm.fr/paradb/	Paralogy mapping in human genomes
TRBase	http://bioinfo.ex.ac.uk/trbase	Human genome tandem repeats
UCSC genome	http://genome.ucsc.edu/	Genome assemblies and annotation
Comparative genomics		
ArkDB	http://www.thearkdb.org/	Farm & other animal genomes
ChickVD	http://chicken.genomics.org.cn/	Seq variation in chicken genome
FANTOM	http://fantom2.gsc.riken.go.jp	Mouse cDNA clone functional annotation
GALA	http://gala.cse.psu.edu/	Genomic align, annotation, exptl results
HomoloGene	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene	Homologous genes in complete euk genomes
Inparanoid	http://inparanoid.cgb.ki.se/	Euk orthologs
PhenomicDB	http://www.phenomicdb.de/	Phenotype comparison of orthologous genes in human and model organisms
RatMap	http://ratmap.org/	Rat genome tools, data
TAED	http://www.bioinfo.no/tools/TAED	Phyl-based tools for comparative genomics
VEGA	http://vega.sanger.ac.uk/	Vertebrate genome annotation
Phytome	http:www.phytome.org	Plant comparative genomics
MBGO	http://mbgd.genome.jp/	Comparative microbial genomes
Eukaryotic genomes		
CandidaDB	http://genolist.pasteur.fr/CandidaDB	*Candida albicans* genomes
Candida Genm	http://www.candidagenome.org/	*Candida albicans* genomes
CYGD	http://mips.gsf.de/proj/yeast	MIPS yeast genome data
PROPHECY	http://prophecy.lundberg.gu.se/	Yeast phenotypic characteristics profiling
SGD	http://www.yeastgenome.org/	*Saccharomyces* genomes
YRC PDR	http://www.yeastrc.org/pdr/	Yeast resource center pub data repository
CADRE	http://www.cadre.man.ac.uk/	*Aspergillus* data repository
MNCDB	http://mips.gsf.de/proj/neurospora/	MIPS *Neurospora crassa* data
C. elegans Proj	http://www.sanger.ac.uk/Projects/C_elegans	*C. elegans* genome sequences
WILMA	http://www.came.sbg.ac.at/wilma/	*C. elegans* annotation
WormBase	http://www.wormbase.org/	*C. elegans* & *C. birggsae* curated genomes
FlyBase	http://flybase.bio.indiana.edu/	*Drosophila* seq & genome info
FlyMine	http://www.flymine.org/	Integrated insect genomic, proteomic data
GadFly	http://fruitfly.org	*Drosophila* genome annotation
BeetleBase	http://www.bioinformatics.ksu.edu/BeetleBase	Beetle *Tribolium castaneum* genomes
NEMBASE	http://www.nematodes.org/	Nematode seq & functional data
SpodoBase	http://bioweb.ensam.inra.fr/spodobase/	Butterfly *Spodoptera frugiperda* genomics

TABLE 15.13 ***continued***

Database	URL	Description
Eukaryotic (plant) genomes		
CropNet	http://ukcorp.net/	Crop plants
GrainGenes	http://graingenes.org	Gene and PT of Barley, oats, rye, wheat
Gramene	http://www.gramene.org	Comparative grass genomics
IRIS	http://www.iris.irri.org/	Intl rice information system
MaizseGDB	http://www.maizegdb.org/	Maize genetics and genomics
MtDB	http://www.medicago.org/MtDB	*Medicago* genomes
Oryzabase	http://www.shigen.nig.ac.jp/rice/oryzabase/	Rice genetics and genomics
PlantGDB	http://www.plantdb.org	Plant genome DB
PoMaMo	https://gabi.rzpd.de/PoMaMo.html	Potato genomes
SGMD	http://psi081.ab.ars.usda.gov/SGMD/default.htm	Soybean genomics and microarray
TropGENEDB	http://tropgenedb.cirad.fr/	Banana, cocoa, sugarcane genomes
TAIR	http://www.arabidopsis.org/	*Arabidopsis* information resource
Prokaryotic genomes		
BacMap	http://igs-server.cnrs-mrs.fr/axenic/	Atlas of annotated bacteria genomes
BioCyc	http://biocyc.org/	Bacterial genomes
ASAP	http://asap.ahabs.wisc.edu/annotation/php/ASAP1.html	Systematic annotation of *E. coli* and related genomes
coliBase	http://colibase.bbam.ac.uk/	*E. coli, Salmonella* and *Shigella* data
EchoBase	http://www.ecoli-york.org/	Post-genomic studies of *E. coli*
EcoGene	http://bmb.med.miami.edu/EcoGene/EcoWeb/	*E. coli* gene and protein seq & literatures
GenProEC	http://genprotec.mbl.edu	*E. coli* K12 genome, proteome data
PEC	http://shigen.lab.nig.ac.jp/ecoli/pec	*E. coli* chromosome profiling
BSORF	http://bacillus.genome.ad.jp/	*Bacillus subtilis* genomes
SubiList	http://genolist.pasteur.fr/SubtiList/	*Bacillus subtilis* genomes
CampyDB	http://campy.bham.ac.uk/	*Campylobacter* genome analysis
ClostriDB	http://clostri.bham.ac.uk/	*Clostridium* spp. Genomes
CIDB	http://www.it.deakin.edu.au/CIDB	*Chlamydia* gene expression data
CyanoBase	http://www.kazusa.or.jp/cyano	Cyanolbacterial genomes
LeptoList	http://bioinfo.hku.hk/LeptoList	*Leptospira interrogans* genomes
VirFact	http://virfact.burnham.org/	Bacterial virulence factors
VFs	http://zdsys.chgb.org.cn/VFs/Main.htm	References for microbial virulence factors
Viral genomes		
HCVDB	http://hepatitis.ibep.fr/	Hepatitis C virus
HIV ResistDB	http://resdb.lanl.gov/Resist_DB/default.htm	HIV mutations resistant to anti-HIV drug
NCBI Viral genome	http://www.ncbi.nlm.nih.gov/genome/VIRUSES/viruses.html	Viral genome resource at NCBI
Poxvirus	http://www.poxvirus.org/	Poxvirus genomic seq, gene annotation

Note: Abbreviations used: align, alignment; euk, eukaryotics; DB, database; genm, genome; intl, international; prok, prokaryotics; FG, functional genomics; GP genome project(s); chrom, chromosome(al); NR, non-redundant; seq, sequence(s); phyl, phylogeny; PT, phenotype(s); pub, public; proj, project.

finished human genome sequence will probably be necessary for the construction of an accurate gene map. Table 15.13 provides a list of web sites where initiatives are ongoing to establish integrated genome databases.

The immediate medical application of human genome information is in identification and annotation of genes associated with disease, the pursuit of new diagnostics, and treatments for these diseases. About 5000 human diseases are known to have a genetic component and 1000 disease-associated markers or

TABLE 15.14 Representative human cancer databases

Database	URL	Description
Atlas, oncology and haematology	http://www.infobiogen.fr/services/chromcancer/	Cancer-related genes, chrom abnorm in oncolgy, haematology and cancer- prone diseases
Cancer chromosome	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi_db=cancerchromosomes	Cytogenetic, clinical and ref info on cancer-related aberrations
CGED	http://love2.aist-nara.ac.jp/CGED	CG expression DB
OncoMine	http://www.oncomine.org/	Cancer MA by gene or cancer type
Oral CG DB	http://tumore-gene.org/Oral/oral.html	Cellular and molecular data for oral CG
Tumar gene family DB	http://www.tumor-gene.org/tgdf.html	Cellular, molecular and biological data for various CG

Note: Abbreviations used: CG, cancer gene(s); MA, microarray, NW, network.

genes have already been isolated over the years (Antonarakis and McKusick, 2000). A disease gene can be localized to a particular chromosome by:

- identifying gross chromosome rearrangements, such as translocations, deletions and duplications, which serve as markers for the position of the gene; or
- by collecting pedigrees in which the responsible gene is segregating and performing linkage analysis or association studies with polymorphism markers.

Determining the genetic component(s) of non-Mendelian characters such as heart disease, asthma and cancer is more difficult. These complex traits can be continuous or discontinuous, and their phenotypic expression often depends on the interaction of a myriad of genetic, social and environmental factors. However, by looking at the degree of disease clustering within families, it may become apparent that a complex disease has a heritable component. The mode of inheritance can be determined from epidemiological data taken from large populations or within affected pedigrees. The HGP and its associated technologies hold great promise in aiding in the analysis and identification of important polymorphisms for every gene (Risch and Merikangas, 1996). Examples of medically relevant genes cloned using HGP information include those involved in inherited breast cancer (Miki *et al.*, 1994), early onset Alzheimer disease (Sherrington *et al.*, 1995), and the hereditary non-polyposis colorectal cancer (Fishel *et al.*, 1993; Papadopoulos *et al.*, 1994). A large number of diseases-related databases are available online. Sample cancer databases are given in Table 15.14.

15.6 REFERENCES

Aaronson, J., Eckman, B., Blevins, R.A. *et al.* (1996) *Genome Research*, **6**: 829–45.

Alexandersson, M., Cawley, S. and Pachter, L. (2003) *Genome Research*, **13**, 496–502.

Antonarakis, S.E. and Mckusick, V.A. (2000) *Nature Genetics*, **1**, 11.

Baekelandt, I.N., Goritchenko, L. and Benowitz, L.I. (1997) *Nucleic Acids Research*, **25**, 1281–8.

Benfey, P.N. and Protopapas, A.D. (2005) *Genomics*, Prentice-Hall, Upper Saddle River, NJ.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J. *et al.* (2006) *Nucleic Acids Research*, **34**, D16–D20.

Boon, K., Osorio, E.C., Greenhut, S.F. *et al.* (2002) *Proceedings of the National Academy Sciences, USA*, **99**, 11287–92.

Bork, P., Dandekar, T., Diaz-Lazoz, Y. *et al.* (1998) *Journal of Molecular Biology*, **283**: 707–25.

Borodovsky, M.Y. and Mcininch, J.D. (1993) *Computer Chemistry*, **17**: 123–33.

Bourne, P.E. and Weissig, H. (eds.) (2003) *Structural Bioinformatics*, John Wiley & Sons, Inc. New York.

Bowtell, D.D.L. (1999) *Nature Genetics*, **21**, 25–32.

Brent, M.R. and Guigó, R. (2004) *Current Opinions in Structural Biology,* **14**, 264–72.

Burge, C. and Karlin, S. (1997) *Journal of Molecular Biology*, **268**, 78–94;

Bourne, P.E. and Weissig, H. (eds.) (2003) *Structural Bioinformatics*, John Wiley & Sons, Inc., New York.

Burge, C. and Karlin, S. (1997) *Journal of Molecular Biology*, **268**, 78–94

Canter, C.R. and Smith, C.L. (1999) *Genomics*, John Wiley & Sons, Inc., Hoboken, NJ.

Claverie, J.-M. (1997) *Human Molecular Genetics*, **6**. 1735–44.

Cowell, I.G. and Austin, C.A. (eds) (1997) *cDNA Library Protocols*, Humana Press, Totowa, NJ.

Derisi, J., Penland, L., Brown, P.O. *et al.* (1996) "Use of cDNA microarray to analyze gene expression in human cancer." *Nature Genetics*, **14**, 457–60.

Emmert, D.B., Stoehr, P.J., Stoesser, G. and Cameron, G.N. (1994) *Nucleic Acids Research*, **22**, 3445–9.

Fickett, J.W. (1996) *Trends in Genetics*, **12**, 316–20.

Fickett, J.W. and Hatzigeorgiou, A.G. (1997) *Genome Research*, **7**, 861–78.

Fickett, J.W. and Tung, C.-S. (1992) *Nucleic Acids Research*, **20**, 6441–50.

Fishel, R., Lescoe, M., Rao, M.R.S. *et al.* (1993) *Cell*, **75**, 1027–38.

Fleischmann, R.D. *et al.* (1995) *Science*, **269**, 496–512.

Fodor, S.P., Read, J.L., Pirrung, L. *et al.* (1991) *Science*, **251**, 767–73.

Foissac, S., Bardou, P., Moisan, A. *et al.* (2003) *Nucleic Acids Research*, **31**, 3742–5.

French, T., Gultyaev, A.P. and Gerdes, K. (1997) *Journal of Molecular Biology*, **273**, 38–51.

Gelfand, M.S. (1995) *Journal of Computer Biology*, **2**, 87–115.

Gelfand, M.S., Mironov, A.A. and Pevzner, P.A. (1996) *Proceedings of the National Academy of Sciences, USA*, **93**, 9061–6.

Green, P., Lipman, D., Hillier, L. *et al.* (1993) *Science*, **229**, 1711–6.

Guigó, R., Knudsen, S., Drake, N., and Smith, T. (1992) *Journal of Molecular Biology*, **226**, 141–57.

Huang, X., Adams, M.D., Zhou, H. and Kerlevage, A. (1997) *Genomics*, **46**, 37–45.

Jayasena, S.D. (1999) *Clinical Chemistry*, **45**, 1628–50.

Johnson, J.M., Castle, J., Garnett-Engele, P. *et al.* (2003) *Science*, **302**, 2141–4.

Kanz, C., Aldebert, P., Althorpe, N. *et al.* (2005) *Nucleic Acids Research*, **33**, D29–33.

Klung, S.J. and Famulock, M. (1994) *Molecular Biology Reports*, **20**, 97–107.

Korf, I., Flicek, P., Duan, D. and Brent, M.R. (2001) *Bioinformatics*, **17**, S140–8;

Kozak, M. (1996) *Mammalian Genome*, **7**, 563–74.

Krogh, A. (1977) *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, **5**, 179–86;

Lee, J.F., Hesselberth, J.R., Meyers, L.A. and Ellington, A.D. (2004) *Nucleic Acids Research*, **32**, D95–D100.

Lee, Y., Tsai, J., Sunkara, S., Karamycheva, S. *et al.* (2005) *Nucleic Acids Research*, **33**, D71–4.

Liolios, K., Tavernarakis, N., Hugenholtz, P. and Kyrpides, N.C. (2006) *Nucleic Acid Resear*, **34**, D332–D334.

Lukashin, A.V. and Borodovsky, M. (1998) *Nucleic Acids Research*, **26**, 1107–15.

McGinnis, S., Madden, T.L. (2004) *Nucleic Acids Research*, **32**, W20–5

McKusick, V.A. (1997) *Genomics*, **45**, 244–9.

Meyer, I.M. and Durbin, R. (2002) *Bioinformatics*, **18**, 1309–18.

Miki, Y., Swensen, J., Shattuck-Eidens, D. *et al.* (1994) *Science*, **266**, 66–71.

Mikilos, G.L.G. and Rubin, G.M. (1996) *Cell*, **86**, 521–9.

Müller, U.R. and Nicolau, D.V. (eds) (2005) *Microarray Technology and its Applications*, Springer, Berlin, Germany.

Murthy, V.L. and Rose, G.D. (2003) *Nucleic Acids Research*, **31**, 502–4.

Papadopoulos, N., Nicolaides, N.C., Wei, Y.-F. *et al.* (1994) *Science*, **263**, 1625–9.

Parra, G., Agarwal, P., Abril, J.F. *et al.* (2003) *Genome Research*, **13**, 108–17

Perou, C.M., Jeffrey, S.S., van de Rijin, M. *et al.* (1999) 'Distinct gene expression patterns in human mammary epithelial cells and breast cancers.' *Proceedings of the National Academy of Sciences, USA,* **96**, 9212–7.

Pesole, G., Liumi, S., Grillo, G. *et al.* (2002) *Nucleic Acids Research*, **30**, 335–40.

Pevsner, J. (2003) *Bioinformatics and Functional Genomics*, Wiley-Liss, Inc. Hoboken, NJ.

Pleasance, E.D., Marra, M.A. and Jones, S.J. (2003) *Genome Research*, **13**, 1203–15.

Ponomarenko, J.V., Orlova, G.V., Ponomarenko, M.P. *et al.* (2000) *Nucleic Acids Research*, **28**, 205–8.

Prashar, Y. and Weissmann, S.M. (1996) *Proceedings of the National Academy of Sciences, USA*. **93**, 659–63.

Praz, V., Périer, R., Bonnard, C. and Bucher, P. (2002) *Nucleic Acid Research*, **3**, 322–4.

Reese, M.G., Eeckman, F.H., Kulp, D. and Haussler, D.J. (1997) *Journal of Computer Biology*, **4**, 311–23.

Renner, C., Trumper, L., Pfitzenmeier, J.-P. *et al.* (1998) *BioTechniques*, **24**, 720–4.

Richmond, C.S., Glasner, J.D., Mau, R. *et al.* (1999) *Nucleic Acids Research*, **27**, 3821–35.

Risch, N. and Merikangas, K. (1996) *Science*, **273**, 1516–7.

Roulet, E., Busso, S., Camargo, A.A. *et al.* (2002) *Nature Biotechnology*, **20**, 831–5.

Saha, S., Sparks, A.B., Rago, C. *et al.* (2002) *Nature Biotechnology*, **20**, 508–12.

Schena, M. (2003) *Microarray Analysis*, John Wiley & Sons, Inc., New York.

Senapathy, P., Shapiro, M.B. and Harris, N.L. (1990) *Methods in Enzymology*, **18**, 252–78.

Senterre-Lesenfants, S., Alag, A.S. and Sobel, M.E. (1995) *Journal of Cell Biochemistry*, **58**, 445–54.

Shapiro, L. and Harris, T. (2000) *Current Opinions in Biotechnology*, **11**, 31–5.

Sherrington, R., Rogaev, E.I., Liang, Y. *et al.* (1995) *Nature*, **375**, 754–60.

Shiue, L. (1997) *Drug Developments Research*, **41**, 142–59.

Singer, M. and Berg, P. (1991) *Genes and Genomes: A Changing Perspective*, University Science Books, Mill Valley, CA.

Smit, A.F.A. (1996) *Cun. Opin. Gen. Dev.*, **6**, 743–9.

Solovyev, V.V., Salamov, A.A. and Lawrence, C.B. (1994) *Nucleic Acids Research*, **22**, 5156–63.

Stanke, M., Steinkamp, R., Waack, S. and Morgenstern, B. (2004) *Nucleic Acids Research*, **32**. W309–12.

Starkey, M.P. and Elaswarapu, R. (eds). (2001) *Genomics Protocols*, Humana Press, Totowa, NJ.

Stekel, D. (2003) *Microarray Bioinformatics*, Cambridge University, Press, Cambridge, UK.

Strachan, T. and Read, A.P. (1999) *Human Molecular Genetics*, BIOScience, Oxford, UK.

Taher, L., Rinner, O., Garg, S. *et al.* (2004) *Nucleic Acids Research*, **32,** W305–8.

Tamura, M., Hendrix, D.K., Klosterman, P.S. *et al.* (2004) *Nucleic Acids Research*, **32**, D182–4.

Tateno, Y., Saitou, N., Okubo, K. *et al.* (2005) *Nucleic Acids Research*, **33**, D25–8.

Uberbacher, E.C. and Mural, R.J. (1991) *Proceedings of the National Academy of Sciences, USA*, **88**, 11261–5.

Vasmatzis, G., Essand, M., Brinkmann, U. *et al.* (1998) *Proceedings of the National Academy of Sciences, USA*, **95**, 300–4.

Velculescu, V.E., Zhang, L., Zhou, W. *et al.* (1997) *Cell*, **88**, 243–51.

Venter, J.C. *et al.* (2001) *Science*, **291**, 1304–51.

Wahle, E. and Keller, W. (1996) *Trends in Biochemical Science*, **21**, 247–50.

Wang, K., Gan, L., Jeffery, E. *et al.* (1999) *Gene*, **229**, 101–8.

Weber, J.L. and Myers, E.W. (1997) *Genome Research*, **7**, 401–9.

Wodiscka, L., Dong, H., Mittmann, M. *et al.* (1997) *Nature Biotechnology,* **15**, 1359–67.

Young, R.A. (2000) *Cell*, **102**, 9–15.

World Wide Webs cited

Aptamer database:	http://aptamer.icmb.utexas.edu/
Array expression database:	http://www.ebi.ac.uk/arrayexpress/
Bibliographies for computational gene recognition	http://linkage.rockefeller.edu/wli/gene/right.html
Cellular Response Database:	http://LH15.umbc.edu/crd
CODEHOP:	http://bioinformatics.weizmann.ac.il/blocks/codehop.html
Codon usage (CUTG):	http://www.kazusa.or.jp/codon/
Codon usage database:	http://biochem.otago.ac.nz:800/Transterm/homepage.html
DbEST:	http://www.ncbi.nlm.nih.gov/dbEST/index.html
Digital Differential Display:	http://www.ncbi.nlm.nih.gov/UniGene/ddd.cgi?ORG=Hs
DNA Data Bank of Japan (DDBJ):	http://ddbj.nig.ac.jp
EMBL Nucleotide Sequence DB at EBI:	http://www.ebi.ac.uk/embl.html
Entrez:	http://www.ncbi.nkm.nih.gov/Entrez
EID:	http://mcb.harvard.edu/gilbert/EID/
EPD:	http://www.epd.isb-sib.ch
ExInt database:	http://intron.bic.nus.edu.sg/exint/extint.html
GenBank (NCBI):	http://www.ncbi.nlm.nih.gov/GenBank
GeneCards:	http://bioinformatics.weizmann.ac.il/cards/
Genetic Codes:	http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi Globin
Gene Server:	http://globin.csc.psu.edu
GOLD:	http://www.genomesonline.org
Human Developmental Anatomy:	http://www.ana.ed.ac.uk/anatomy/database/
Intron:	http://nutmeg.bio.indiana.edu/intron/index.html
NDB:	http://ndbserver.rutgers.edu/
NEBcutter:	http://tools.neb.com/NEBcutter2/index.php
PDA:	http://dbb.nhri.org.tw/primer/
PDB:	http://www.rcsb.org/pdb
PlantProm:	http://mendel.cs.rhul.ac.uk/
RepeatMasker:	http://www.genome.washington.edu/analysistools/repeatmask.htm
RNABase:	http://www.rnabase.org/
SAGEmap:	http://www.ncbi.nlm.nih.gov/SAGE
SAGEnet:	http://www.sagenet.org
SAGEGenie:	http://cgap.nci.nih.gov/SAGE
SCOR:	http://scor.lbl.gov/
SEQANALREF at ExPASy:	http://expasy.hcuge.ch
SELEX_DB:	http://wwwmgs.bionet.nsc.ru/mgs/systems/selex/
SRS retrieval system:	http://srs.ebi.ac.uk/

TIGR Gene Indices of TDB:	http://www.tigr.org/tdb/tdb.html
Transterm:	http://uther.otago.ac.nz/Transterm.html
UniGene:	http://www/ncbi.nlm.nih.gov/UniGene/
UTRdb:	http://bigrea.area.ba.cnr.it:8000/srs6/
WIGS:	http://www.genome.wi.mit.edu/
Webcutter:	http://www.firstmarket.com/cutter/cut2.html
XProfiler:	http://www.ncbi.nlm.nih.gov/ncicgap/cgapxpsetup.cgi
Useful genome databases:	Table 15.2
EST databases:	Table 15.3
Alternative splicing databases:	Table 15.4
Single nucleotide polymorphisam databases:	Table 15.6
Various RNA sequence databases:	Table 15.7
Nucleic acid secondary databases:	Table 15.8
Computational genomic serves:	Table 15.9
Gene identification servers:	Table 15.10
Microarray databases and servers:	Table 15.12
Genome project databases:	Table 15.13
Human cancer databases:	Table 15.14

CHAPTER 16

PROTEOMICS

16.1 PROTEOME: FEATURES AND PROPERTIES

16.1.1 Proteome features

The progress in genome sequencing has been dramatic. The completed genomes provide unprecedented opportunity for advancing our understanding biochemistry and molecular biology of organisms (Augen, 2004). The integration of sequence information with differential gene expression and function data produces enormous quantities of information, known as functional genomics. Ultimately the analysis of the expressed level, localization, structure and function of the protein products of the genome will define the activity of a cell or organism, thus the importance of proteomics to support genome-scale analysis of protein structures, interactions and functions.

A proteome can be considered as the comprehensive group of proteins expressed by a given cell or tissue. Proteomics can then be defined as the systematic analysis and documentation of the proteins in biological systems. It is the analysis of the protein complement (proteome) expressed by a genome or a cell or a tissue (Blackstock and Weir 1999; Liebler, 2002; Zhu *et al.*, 2003). Proteomics can be visualized as a biochemical screening approach, which aims to document the overall distribution of proteins in cells, identify and characterize individual proteins of interest, and ultimately to elucidate their relationships and functional roles. Accordingly, proteomics can be divided into expression proteomics that studies global changes in protein expression and cell-map proteomics, which systematically investigate protein-protein interactions (Waksman, 2005). Distinctions in protein expression are characteristics of different cell types and function and of phenotypic differences within a given cell type.

Although a proteome is defined as the total complement of a genome or the set of proteins encoded by a genome, the strict definition which describes the proteome as the protein readout of the genome is inadequate because:

- the proteome are dynamic as not all proteins are expressed at the same time in a particular cell, tissue or organism,
- the genome does not explicitly encode the full structure and diversity of the proteins, such as alternative splicing and posttranslational modifications (e.g. different glycoforms of glycoproteins via posttranslational glycosylation) expanding proteome diversity in an organism.

Unlike the genome, which is complete in its entirety in almost all cells, the proteome is highly cell specific. Different cells express different subsets of the total protein complement of an organism. Similarly the genome is a static repository of information whose content do not change with time, whereas the proteome is highly dynamic with the subset of expressed proteins changes with time according to the development and physiological

---

Biomacromolecules, by C. Stan Tsai
Copyright © 2007 John Wiley & Sons, Inc.

state of the cell. Once produced, the protein can potentially undergo posttranslational modifications before assuming its role in cell activity. Since its structural modifications and accompanied changes are not all controlled by the same gene, the active form of the individual proteins cannot be determined with reference to any single gene. Furthermore, posttranslational modifications are known to be affected by developmental and physiological states of the cell. Therefore even if all the genes of a genome have been sequenced and translated into proteins, it remains a task to identify and characterize the various functional forms of each protein in a particular cell type under certain conditions. These features of proteome present a major challenge to systematic analysis and documentation of the proteome of a multicellular organism, because the protein profile of each cell type in each of its functional forms in possible developmental and physiological states needs to be specified/annotated. Thus the primary goal of proteomics is the understanding of the structure, function, expression, cellular localization, interacting partners and regulation of every protein produced from a complete genome.

Expression proteomics studies the differences in expressed proteins and regulation of protein levels, while functional proteomics aims at the characterization of the protein complement in cellular compartments, functions, interactions and signal transductions. Approaches to protein expression analysis fall broadly into two categories (Westermeier and Naven, 2002); separation-dependent such as two-dimensional electrophoresis in combination with mass spectrometry and separation-independent such as protein chips (microarrays). Structural and functional analyses of the expressed proteins are essential steps toward understanding biological processes (Lundblad, 2005). These include protein-structure analysis (Swindells *et al.*, 1998; Tsigelny, 2002; Webster, 2000; Wery and Schevitz. 1997), prediction of protein function (Pellegrini *et al.*, 1999; Pennington and Dunn, 2001), mapping protein–protein interactions (Uetz *et al.*, 2000; Waksman, 2005) and computational method (Brooks *et al.*, 1988; Enright *et al.*, 1999; Tsai, 2002). Table 16.1 gives some general proteome resources.

16.1.2 Protein identity based on composition and properties

The proteome, unlike the genome, is a dynamic entity because it is the product of both gene expression and posttranslational alternations, and varies with different cell types, different stages of development and is greatly affected by the environment. Therefore the first step in the proteomic research lies in optimizing the methods used to analyze and identify the individual proteins.

Polyacrylamide gel electrophoresis (PAGE) and high performance liquid chromatography (HPLC), which separate, detect, and quantify proteins present in a given system in a manner that also measures the protein's molecular weight, isoelectric point (p*I*) and other properties, are useful techniques in proteomic research.

A number of useful computational tools have been developed for predicting the identity of unknown proteins based on the physical and chemical properties of amino acids and *vice versa*. Many of these tools are available through the Expert Protein Analysis System (ExPASy) at http://www.expasy.org (Gasteiger *et al.*, 2003) and other servers.

Rather than using an amino acid sequence to search SWISS-PROT, AACompIdent of ExPASy Proteomic tools (http://www.expasy.org/tools/) uses the amino acid composition of an unknown protein to identify known proteins of the same composition. The program requires the desired amino acid composition, the pI and molecular weight of the protein (if known), the appropriate taxonomic class and any special keywords. The user must select from one of six amino acid constellations that influence how the analysis is performed. For each sequence in the database, the algorithm computes a score based on

TABLE 16.1 Useful general proteome databases

Database	URL	Description
3D-GENOMICS	http://www.sbg.bio.ic.ac.uk/3dgenomics	Proteome structural annotations
ExPASy	http://www.expasy.org/	Integrated analytical tools. links
ExProt	http://www.cmbi.kun.nl/EXProt/	Protein sequences verified function
IMB Jena Image Library	http://www.imb-jena.de/IMAGE.html	Visualization and analysis of 3D biomacromolecular structures
iProClass	http://pir.georgetown.edu/iproclass	Integrated protein classification
ISSD	http://www.protein.bio.msu.su/issd	Integrated sequence-structure DB
ModBase	http://alto.compbio.ucsf.edu/modbase-cgi/index.cgi	Annotated comparative protein structures
NegProt	http://superfly.ucsd.edu/negprot	Complete proteome comparison
NCBI ProteinDB	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein	All protein sequences
PDB	http://rcsb.org/pdb	3D structures of proteins, nucleic acids
SWISS-MODEL Repository	http://swissmodel.expasy.org/repository	Annotated 3D protein structure models
UniProt	http://www.uniprot.org/	Universal protein knowledgebase merged data from PIR, Swiss-Prot and TrEMBL
KEGG	http://www.genome.jp/kegg	Integrated DB on genes, proteins, metabolic pathways
PathDB	http://www.ncgr.org/pathdb	Biochem pathways, compds, metabolism
Reactome	http://www.reactome.org/	Knowledgebase of biological pathways

the difference in compositions between the sequence and the query composition. The results, returned by e-mail, are organized as three ranked lists. Because the computed scores are a measure of difference, a score of zero implies that there is exact correspondence between the query composition and that sequence entry. AACompSim (ExPASy Proteomic tools) performs a similar type of analysis, but rather than using an experimentally derived amino acid composition as the basis of searches, the sequence of a SWISS-PROT protein is used instead. A theoretical pI and molecular weight are computed prior to computing the difference score using Compute pI/MW.

PROPSEARCH (http://www.embl-heidelberg.de/prs.html) is designed to find the putative protein family if alignment methods fail (Hobohm and Sander, 1995). The program uses the amino acid composition of a protein taken from the query sequence to discern members of the same protein family by weighing 144 different physical properties including molecular weight, the content of bulky residues, average hydrophobicity, average charge, etc. in performing the analysis. This collection of physical properties is called the query vector, and is compared against the same type of vector pre-computed for every sequence in the target databases (SWISS-PROT and PIR). The search results are returned by e-mail.

16.1.3 Physicochemical properties based on sequence

Compute pI/MW (ExPASy Proteomic tools) is a tool that calculates the isoelectric point and molecular weight of an input sequence. Molecular weights are calculated by the addition of the average isotopic mass of each amino acid in the sequence plus that of one water molecule. The sequence can be furnished by the user in FASTA format, or by a SWISS-PROT identifier. If a sequence is furnished, the tool automatically computes the pI and molecular weight for the entire length of the sequence. If a SWISS-PROT identifier is given, the user may specify a range of amino acids so that the computation is done on a fragment rather than on the entire protein. The absorption coefficient and pI value are also calculated at aBi (http://www.up.univ-mrs.fr/~wabim/d_abim/compo-p.html). If ‘courbe de titrage’ is checked, the titration curve based on the query sequence is returned.

PeptideMass (ExPASy Proteomic tools), which is designed for use in peptide mapping experiments, determines the cleavage products of a protein after exposure to a specific protease or chemical reagent. The enzymes and reagents available for cleavage via PeptideMass are trypsin, chymotrypsin, Lys C, cyanogen bromide, Arg C, Asp N and Glu C. Theoretically one unique peptides would be sufficiently unambiguous to identify each parent protein if such unique peptides are present and can be isolated for identification. An alternative approach to the unique peptide profiling is to target and isolate peptides containing unique (N- or C-terminal) or rare (C, H, M, Y) amino acids in so called amino acid-based profiling. The effect of the amino acid-based profiling on the human proteome is shown in Table 16.2.

The AAindex database at http://www.genome.ad.jp/aaindex/ (Kawashima and Kanehisa, 2000) collects physicochemical properties of amino acids such as molecular weight, bulkiness, polarity, hydrophobicity, average area buried/exposed, solvent accessibility and secondary structure parameters (propensities for amino acid residues to form α-helix, β-strand, reverse turn and coil structures of proteins). The physicochemical properties can be plotted along the sequence using ProtScale at ExPASy (http://www.expasy.org/cgi-bin/protscale.pl). The hydrophobicity plot can reveal the membrane-associated regions of proteins. TMpred of EMBNet at http://www.ch.embnet.org/software/TMPRED_form.html and TMAP (Milpetz *et al.*, 1995) at http://www.embl-heidelberg.de/tmap/tmap_info.html predicts possible transmembrane regions and their topology from the amino acid sequence.

ProTherm (http://gibk26.bse.kyutech.ac.jp/jouhou/protherm/protherm.html) provides comparative thermodynamic data for wild-type and mutant proteins. The same institute also offers an online thermodynamic data on protein-nucleic acid interactions accessible at http://gibk26.bse.kyutech.ac.jp/jouhou/pronit/pronit.html.

TABLE 16.2 The calculated number of tryptic peptides containing the target amino acids

	Total number of peptide	% of peptides
Total tryptic peptides	892 584	100.0
N- or C-terminal peptides	52 816	5.9
Cys containing peptides	237 111	26.6
His containing peptides	274 576	30.8
Met containing peptides	227 773	25.5
Trp containing peptides	157 538	17.6
Cys & His containing peptides	95 238	10.7

Notes: 1. Taken from Zhang *et al*. (2004).

2. Based on International Protein Index (IPI) human sequence database *via* ftp://ftp.ebi.ac.uk/pub/databases/IPI/current/ipi.HUMAN.fasta.gz which contains 52 816 protein entries.

3. The trypic peptides are those ended with Arg or Lys but not followed by Pro.

16.2 PROTEOME INFORMATICS: SEQUENCE DATABASES AND SERVERS

16.2.1 Amino acid sequence

The essence of sequence analysis is the detection of homologous sequences by means of routine database searches, usually with unknown or uncharacterized query sequences (Aitken, 1990). The identification of such relationships is relatively easy when the level of similarity is higher than 50%. If two sequences share less than 20% identity, it becomes difficult or impossible to establish whether they might have arisen via divergence or convergence. Homology is not a measure of similarity but a statement that sequences have a divergent rather than a convergent relationship. Therefore sequences are homologous if they are related by divergence from a common ancestor. Analogy then refers to protein structures that share similar folds but have no demonstrable sequence similarity or proteins that share groups of catalytic residues with almost exactly equivalent spatial geometry, but otherwise have neither sequence nor structural similarity. For example, the His-Asp-Ser catalytic triad of the serine proteases is observed both in subtilisin with a 3-layer $\alpha\beta\alpha$ sandwich structure and chymotrypsin with a 2-domain β-barrel protein. In the latter case, where sequence and structure are different, we can confidently infer that the catalytic sites result from convergent evolution. In the former case, however, where folds are similar but the sequences differ while such folds are usually considered to be analogous, it is sometimes difficult to rule out the existence of a common ancestor (i.e. homology) because structures are more highly conserved than are sequences.

If two proteins have over 45% identical residues in their optimal alignment, the proteins will have similar structures and are likely to have a common or at least similar function. If they have over 25% identical residues, they are likely to have a similar general folding pattern. The region of 18–25% sequence identity is known as the twilight zone in which the sequence homology is marginal and unreliable. Below the twilight zone is a region where pair-wise sequence alignments are uncertain and insignificant. However, lack of significant sequence similarity does not preclude similarity of structures.

It is much easier and quicker to produce sequence information than to determine 3D structures of proteins in atomic detail. As a consequence there is a protein sequence/structure deficit. In order to benefit from the wealth of sequence information, we must establish, maintain and disseminate sequence databases; provide user-friendly software to access the information, and design analytical tools to visualize and interpret the structural/functional clues associated with these data (Higgins and Taylor, 2000; Pennington and Dunn, 2001).

In the context of protein sequence informatics, we will encounter primary, secondary and boutique databases. Primary and secondary databases are used to address different aspects of sequence analysis. Boutique databases are specialized resources serving selected interest groups. The primary structure (amino acid sequence) of a protein is stored in primary databases as linear alphabets that represent the constituent residues. The secondary structure of a protein corresponding to region of local regularity (e.g., α-helices, β-strands and turns), which in sequence alignments are often apparent as conserved motifs, is stored in secondary databases as patterns. The tertiary structure of a protein derived from the packing of its secondary structural elements, which may form folds and domains, is stored in structure databases as sets of atomic co-ordinates. The protein–protein interactions, which are generally related to protein functions and regulation, are grouped with biochemical function/profiling databases.

16.2.2 Primary sequence database

Some of the important protein sequence databases are:

PIR (Protein Information Resource)-PSD:	http://pir.georgetown.edu/
SWISS-PROT (at ExPASy):	http://www.expasy.org/sprot
SWISS-PROT (at EBI):	http://www.ebi.ac.uk/
MIPS (Munich Info Center for Protein Sequences):	http://www.mips.biochem.mpg.de/
TrEMBL:	http://www.expasy.org/sprot
UniProt:	http://www.uniprot.org/

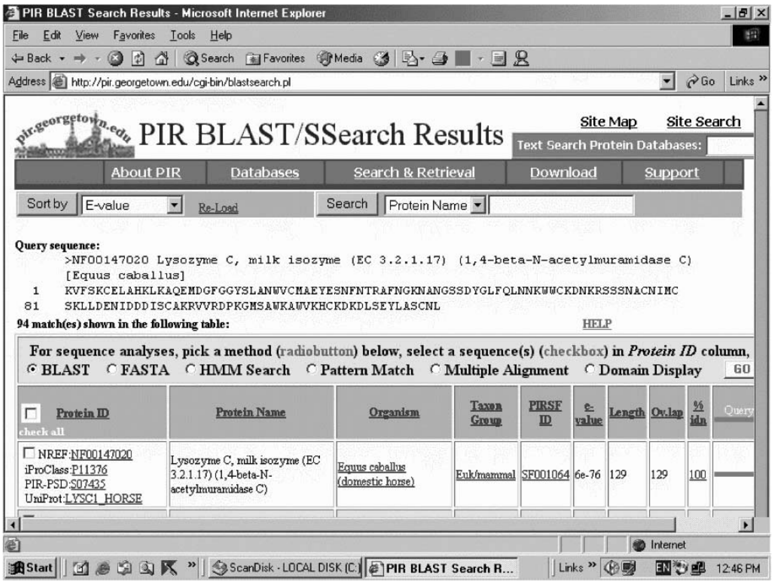
The PIR Protein Sequence Database (Barker *et al.*, 2001; Wu *et al.* 2002) developed at the National Biomedical Research Foundation (NBRF) has been maintained by PIR-International Protein Sequence Database (PSD), which is the largest publicly distributed and freely available protein sequence database. The consortium includes PIR at the NBRF, MIPS and JIPID. PIR-International provides online access at http://pir.georgetown.edu to numerous sequence and auxiliary databases. These include PSD (annotated and classified protein sequences), iProClass (an integrated protein classification database), ASDB (similarity database), NRL3D (sequences from 3D structure database of PDB), RESID (post-translational modification database), and ALN (PIR-alignment database). The PIR server offers three types of database searches:

1. interactive text-based search that allow Boolean queries of text field;
2. standard sequence similarity search including Peptide Match, Pattern Match, BLAST, FASTA, Pair-wise Alignment and Multiple Alignment; and
3. advanced search that combines sequence similarity and annotation searches or evaluate gene family relationships.

Each PIR entry consists of Entry (entry ID), Title, Alternate_names, Organism, Date, Accession (accession number), Reference, Function (description of protein function), Comment (e.g. enzyme specificity and reaction, etc), Classification (superfamily), Keywords (e.g. dimer, alcohol metabolism, metalloprotein, etc), Feature (lists of sequence positions for disulfide bonds, active site and binding site amino acid residues, etc), Summary (number of amino acids and the molecular weight) and Sequence (in PIR format, Chapter 4). In addition, links to PDB, KEGG, BRENDA, WIT, alignments and iProClass are provided.

The PIR site also offers facilities for sequence similarity search (BLAST or FASTA), alignment (ClustalW), and analysis (ProClass, ProtFam, RESID, SSEARCH) of proteins. To perform the similarity search, select BLAST or FASTA to open the search form. Paste the query sequence and click the Submit button. The search returns a list of hits with option(s) for modifying the entries (modify query by choosing refine, add to or remove from the list) followed by pair-wise alignment of query sequence versus retrieved sequences (Figure 16.1). The Modify option(s) enable the user to design the list of desired entries.

SWISS-PROT is a curated protein sequence database (Bairoch and Apweiler, 2000) which, from its inception in 1986, was produced collaboratively by the Department of Medical Biochemistry at the University of Geneva and the EMBL. The database is now maintained collaboratively by Swiss Institute of Bioinformatics (SIB) and EBI/EMBL and provides high-level annotations, including descriptions of the function of the protein and


PIR BLAST Search Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://pir.georgetown.edu/cgi-bin/blastsearch.pl
Go Links
pir.georgetown.edu
PIR BLAST/SSearch Results
Site Map
Site Search
Text Search Protein Databases:
About PIR
Databases
Search & Retrieval
Download
Support
Sort by
E-value
Re-Load
Search
Protein Name
HELP
60
Internet
Start
ScanDisk - LOCAL DISK (C:)
PIR BLAST Search R...
Links
12:46 PM

Query sequence:

```
     >NF00147020 Lysozyme C, milk isozyme (EC 3.2.1.17) (1,4-beta-N-acetylmuramidase C)
     [Equus caballus]
  1  KVFSKCELAHKLKAQEMDGFGGYSLANWVCMAEYESNFNTRAFNGKNANGSSDYGLFQLNNKWWCKDNKRSSSNACNIMC
 81  SKLLDENIDDDISCAKRVVRDPKGMSAWKAWVKHCKDKDLSEYLASCNL
```

94 match(es) shown in the following table:

For sequence analyses, pick a method (radiobutton) below, select a sequence(s) (checkbox) in *Protein ID* column,

● BLAST ○ FASTA ○ HMM Search ○ Pattern Match ○ Multiple Alignment ○ Domain Display

☐ Protein ID check all	Protein Name	Organism	Taxon Group	PIRSF ID	e-value	Length	Ov.lap	% idn	Query
☐ NREF:NF00147020 iProClass:P11376 PIR-PSD:S07435 UniProt:LYSC1_HORSE	Lysozyme C, milk isozyme (EC 3.2.1.17) (1,4-beta-N-acetylmuramidase C)	Equus caballus (domestic horse)	Euk/mammal	SF001064	6e-76	129	129	100	

Figure 16.1 BLAST return of PIR
The BLAST search of milk lysozyme C (NF00147020) returns a list of hits (with % identity). The list is followed by pair-wise alignment of query sequence *versus* matched sequences.

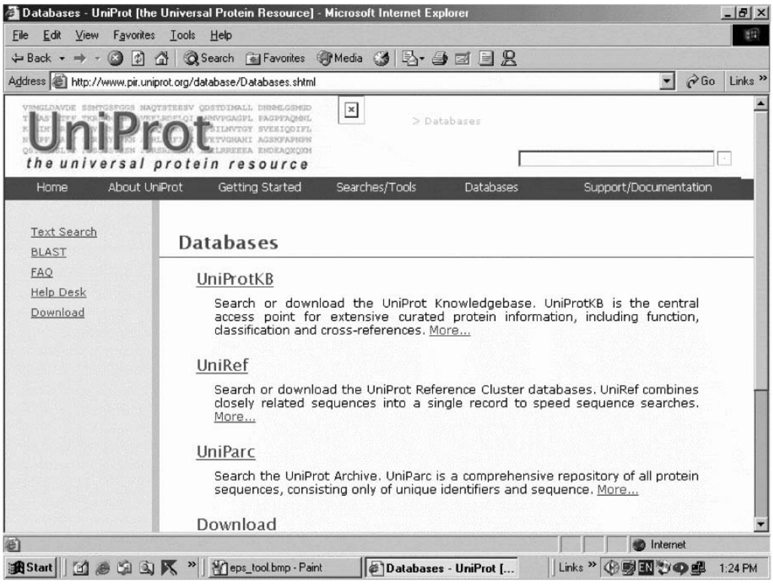
of the structure of its domains, its post-translational modifications, variants and so on. The database can be accessed at http://www.expasy.org or numerous mirror sites. Entries begin with an identification (ID) line and end with a // terminator. Accession number is provided on the AC line. If several numbers appear on the same AC line, the first accession number is the most current. The following lines give the date of entry (DT), the descriptions (DE) the gene name (GN), the organism species (OS), and organism classification (OC). The comment (CC) lines tell us about the FUNCTION of the proteins, its post-translational modifications (PTM), its TISSUE SPECIFICITY, SUBCELLULAR LOCATION, and so on. The CC lines also indicate any known SIMILARITY or affiliation to particular protein families, if such information is available. Following the comment section, database cross-reference (DR) lines list links to other biomolecular databases including primary sources, secondary databases, specialist databases (EMBL, PIR, PDB, Pfam, ProSite, ProDom, ProtoMap, etc). Directly after DR lines are a number of relevant keywords (KW), and a number of Feature Table (FT) lines. The feature table highlights regions of interest in the sequence such as local secondary structure, domains, ligand binding sites, post-translational modifications, etc. The final section of the database entry includes the sequence itself in the single-letter amino acid code on the SQ lines. Sequence data correspond to the precursor form of the protein before post-translational processing and the extent of mature proteins may be deduced by reference to the FT lines, which will indicate the region of a sequence that corresponds to the signal (SIGNAL), transit (TRANSIT) or pro-peptide (PROPEP) respectively.

Translated EMBL (TrEMBL) was created as a computer-annotated supplement to SWISS-PROT (Bleasby *et al.*, 1994) having two main sections. SP-TrEMBL contains entries that are eventually incorporated into SWISS-PROT but not yet been manually annotated. REM-TrEMBL contains sequences that are not destined to be included in SWISS-PROT, including T-cell receptors, fragments of fewer than eight amino acids, synthetic sequences, patented sequences and codon translations that do not encode real proteins.

The amino acid sequences can be searched and retrieved from the integrated retrieval sites such as Entrez (Schuler, *et al.*, 1996), and SRS of EBI (http://srs.ebi.ac.uk/) and DDBJ (http://srs.ddbj.nig.ac.jp/index-e.html). From the Entrez home page (http://www.ncbi.nlm.nih.gov/Entrez), select Protein to open the protein search page to retrieve amino acid sequences of proteins in two formats; GenPept and FASTA. The GenPept format is similar to the GenBank format with annotated information, reference(s) and features. The amino acid sequences of the EBI are derived from the SWISS-PROT database. The retrieval system of the DDBJ consists of PIR, SWISS-PROT and DAD which returns sequences in the GenPept format.

In 2002, UniProt consortium (http://www.uniprot.org) was formed by uniting the SWISS-PROT + TrEMBL and PIR-PSD activities by maintaining a high-quality database that serves as a stable, comprehensive, fully classified and accurately annotated protein sequence knowledge base (Figure 16.2). The database offers extensive cross-references and querying interfaces fully accessible to the scientific community (Bairoch *et al.*, 2005). The UniProt consortium produces three layers of protein sequence databases:

1. The UniProt Archive (UniParc) provides a stable, comprehensive, nonredundant sequence collection by storing the complete body of publicly available protein sequence data. Although most protein sequence data are derived from the translation of DDBJ/EMBL/GenBank sequences, primary protein sequence data are also submitted directly to UniProt or derived from the PDB entries. The Archive also captures protein sequence data from other sources such as Ensemble, International Protein Index (IPI), NCBI-RefSeq, FlyBase, and WormBase. Each protein sequence is assigned to a unique UniParc identifier (UPI#) and represented only once in the Archive. In UniParc, the


Databases - UniProt [the Universal Protein Resource] - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.pir.uniprot.org/database/Databases.shtml
Go Links
UniProt
the universal protein resource
> Databases
Home
About UniProt
Getting Started
Searches/Tools
Databases
Support/Documentation
Text Search
BLAST
FAQ
Help Desk
Download
Internet
Start
eps_tool.bmp - Paint
Databases - UniProt [...
Links
1:24 PM

Databases

UniProtKB

Search or download the UniProt Knowledgebase. UniProtKB is the central access point for extensive curated protein information, including function, classification and cross-references. More...

UniRef

Search or download the UniProt Reference Cluster databases. UniRef combines closely related sequences into a single record to speed sequence searches. More...

UniParc

Search the UniProt Archive. UniParc is a comprehensive repository of all protein sequences, consisting only of unique identifiers and sequence. More...

Download

Figure 16.2 Database page of UniProt
The databases of UniProt are accessible at http://www.uniprot.org/detabase/Databases.shtml which offers searches for UniProtKB, UniRef, UniParc (a comprehensive repository of all sequences) and a download facility.

nonredundancy is taken that all sequences that are 100% identical over their entire length are merged into a single entry, regardless of species. The cross-reference to the source databases uses ‘active’ or ‘obsolete’ to indicate the presence or absence of the entry in the source databases.

2. The UniProt Knowledgebase (UniProtKB) merges PIR-PSD, SWISS-PROT + TrEMBL to provide the central database of protein sequences with accurate and consistent annotations and functional information. Bi-directional cross-references are created for easy tracking of the entries. The UniProt Knowledgebase has two parts: a section of fully, manually annotated records resulting from literature information extraction and curator-evaluated computational analysis, and a section with computationally analyzed records awaiting full manual annotation. Automatic classification and annotation provide automatic large-scale functional characterization and annotation. InterProt classification recognizes domains and classifies all the protein sequences into families and superfamilies. InterPro (http://www.uniprot.org/interpro) is an integrated resource of protein families, domains and sites that amalgamates the efforts of the member databases (Pfam. PROSITE, PRINTS, ProDom, SMART, PIRSF, Superfamily and TIGRFAMs). Extensive cross-references to external data collections are provided. An integration with structural databases is achieved by residue level mapping of sequences from the PDB entries. To minimize database redundancy, differences between sequencing reports due to splice variants, polymorphisms, disease-causing mutations, experimental sequence modifications or simply sequencing errors are indicated in the feature table. In this respect, the definition of nonredundancy in UniProt differs from UniParc in that UniProtKB aims to describe in a single record all protein products derived from a certain gene from a certain species. The database not only gives the whole record an accession number but also assigns to each protein form derived by alternative splicing, proteolytic cleavage and post-translational modification Isoform identifiers.

3. The UniProt Reference databases (UniRef) provide nonredundant data collections based on the UniProtKB and UniParc, in order to obtain complete coverage of sequence space at several resolutions. UniRef databases (sequence collections clustered by sequence identity, for performing faster homology searches) are created as representative protein sequence databases with high information content.

The most efficient and user-friendly way to browse the UniProt databases is via the UniProt website at http://www.uniprot.org, which serves as a portal to the entire UniProt project. It provides database query and data-mining mechanisms, file download capabilities, user support/communication and links to consortium resources.

16.2.3 Secondary sequence database

There are many secondary sequence databases that contain the biologically significant information of sequence analyses (e.g. patterns, motifs and functional sites) from the primary sources. Some representative secondary sequence databases are given in Table 16.3. Because there are several different primary databases and a variety of ways to analyze protein sequences, the type of information stored in each of the secondary databases is different. However, the creation of most secondary databases is based on a common principle that multiple alignments may identify homologous sequences within which conserved regions or motifs may reflect some vital information crucial to the structure or function of these homologous proteins. Motifs have been exploited in various ways to build diagnostic patterns for particular protein families. If the structure and function of the family are known, searches of pattern databases may offer a fast track to the inference of

TABLE 16.3 Protein sequence secondary databases

Database	URL	Description
Conserved residues, motifs and domains		
Blocks	http://blocks.fhcrc.org/	Conserved regions in prot families
eBLOCKS	http://fold.stanford.edu/eblocks/acsearch.html	Highly conserved seq blocks
CDD	http://www.ncbi.nlm.nih.gov/Structure/cdd/	Conserved domain DB
eMOTIF	http://motif.stanford.edu/emotif	Determination/search of seq motifs
Hits	http://hits.isb-sib.ch/	Protein domains and motifs
Knottins	http://knottin.cbs.cnrs.fr	Small prot with disulfide knot
PANAL	http://mgd.ahc.umn.edu/panal/run_panal.html	Composite blocks, patterns, fam
PROSITE	http://ww.expasy.org/prosite	Funct/signature patterns and profiles
ProTeu	http://www.proteus.cs.huji.ac.il/	Signature seq at N- and C-termini
SBASE	http://www.icgeb.org/sbase	Domain sequences and tools
SMART	http://smart.embl.de/	Identification/annotation of domains
Functional/PTM sites		
ASC	http://bioinformatica.isa.cnr.it/ASC/	Biologically active peptides
CSA	http://www.ebi.ac.uk/thronton-srv/databases/CSA/	Active sites, catalytic residues of enzymes with known 3D
COMe	http://www.ebi.ac.uk/come	Metal coordination sites, classif of bioinorganic proteins
Metalloprot DB	http://metallo.scripps.edu/	Metal-binding sites in metalloprot
O-GlycBase	http://www.cbs.dtu.dk/dababases/OGLYCBASE/	O- and C-linked glycosylation sites
Phospho.ELM	http://phospho.elm.eu.org/	S/T/Y Phosphorylation sites
Classification		
ADDA	http://ekhidna.biocenter.helsinki.fi:8080/examples/servlets/adda/	Prot domain classif and singleton
CHOP	http://cubic.bioc.columbia.edu/services/CHOP	Clusters include singleton clusters
Pfam	http://www.sanger.ac.uk/Software/Pfam/	Protein families
PIRSF	http://pir.georgetown.edu/pirsf/	Family/superfamily classification
PRINTS	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/	Hierarchical genes family FP
ProDom	http://www.toulouse.inra.fr/prodom.html	Prot domain families
ProtoMap	http://protomap.cornell.edu/	Hierarchical classif of SP proteins
PANTHER	http://panther.appliedbiosystems.com	Protein fam, funct related subfam
SMART	http://smart.embl-heidelberg.de	Protein domains
SYSYERS	http://systers.molgen.mpg.de/	Partition of fam and superfam
TigrFam	http://www.tigr.org/TIGRFAMs/index.shtml	Protein families

Notes: 1. Abbreviations used; biol, biologically; classif, classification; funct, functional; interact, interaction/interacting; prot, protein(s); seq, sequence(s); FP, fingerprints; PTM, posttranslational modification; SP, Swiss-Prot; fam, families; superfam, superfamilies; funct. Function(al/ally).

2. Singleton refers to the sequence having no relative either within its own genome or any other genome.

biological function of the unknown. Notwithstanding, the pattern databases should be used to augment primary database searches.

In the BLOCKS database (Henikoff *et al.*, 1998), the motifs or blocks are created by automatically detecting the most highly conserved regions of each protein family. Initially three conserved amino acids are identified. The resulting blocks are then calibrated against SWISS-PROT to obtain a measure of the likelihood of a chance match. The database is accessible by keyword and sequence searches via the BLOCKS Web server at the Fred Hitchinson Cancer Research Center (http://www.blocks.fhcrc.org/). The Block includes the SWISS-PROT IDs of the constituent sequences, the start position of the fragment, the sequence of the fragment and a score or weight indicating a measure of the closeness of the relationship of that sequence to others in the block (100 being the most distant).

PROSITE (Hofmann *et al.*, 1999), which uses SWISS-PROT, is maintained collaboratively at the Swiss Institute of Bioinformatics (http://expasy.hcuge.ch/sprot/prosite.html) and Swiss Institute for Experimental Cancer Research (http://www.isrec.isb-sib.ch/). It is rationalized that protein families could be simply and effectively characterized by the single most conserved signature (motif/pattern/profile) observable in a multiple alignment of known homologous proteins. Such signatures usually associate with key biological functions. The database is the collection of sequence information for biologically significant signatures, patterns, motifs, profiles and fingerprints, each of which is associated with an accession number. Some of the descriptors of the PROSITE are:

- post-translational modification sites;
- domain profiles/sequences;
- DNA or RNA associated protein profiles/signatures;
- enzyme active sites/signatures;
- electron transport protein signatures;
- structural protein signatures;
- cytokine and growth factor signatures;
- hormone and active peptide signatures;
- toxin signatures;
- inhibitor signatures;
- secretive protein; and
- chaperone signatures.

Entries are deposited in PROSITE in two files, the data file and the documentation file. In the data file, each entry contains an identifier (ID) and an accession number (AC). A title or description of the family is contained in the DE line, and the pattern itself resides on PA lines. The following NR lines provide technical details about the derivation and diagnostic performance of the pattern. Large numbers of false-positives and false-negatives indicate poorly performing patterns. The comment (CC) lines furnish information on the taxonomic range of the family, the maximum number of observed repeat, functional site annotations, etc. The following DR lines list the accession number and SWISS-PROT identification codes of all the true matches to the pattern (denoted by T) and any possible matches (denoted by P). The last DO line points to the associated documentation file. The documentation file begins with the accession number and cross-reference identifier of its data file. This is followed by a free format description of the family including details of the pattern, the biological role of selected motif(s) if known and appropriate bibliographic references. The on-line ProtSite search can be conducted at ScanProsite tool (selecting Scan a sequence for the occurrence of PROSITE patterns) of ExPASy Proteomic tools (http://www.expasy.org/tools/scnpsite.html), or PPSearch of EBI (http://www2.ebi.ac.uk/ppsearch/) or ProfileScan server (http://www.isrec.isb-sib.ch/software/PFSCAN_form.html).

Most protein families are characterized by several conserved motifs. The PRINTS fingerprint database was developed to use multiple conserved motifs to build diagnostic signatures of family membership (Attwood *et al.*, 1998). If a query sequence fails to match all the motifs in a given fingerprint, the pattern of matches formed by the remaining motifs allows the user to make a reasonable diagnosis. The PRINTS can be accessed by keyword and sequence searches at http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/. There are three sections in a PRINTS entry. In the first section, each fingerprint is given an identifying code, an accession number, number of motifs in the fingerprint, a list of cross-linked databases and bibliographic information. In the second section, the diagnostic performance is given in the summary and the table of results. The last section lists seed motifs and final motifs that result from the iterative database scanning. Each motif is identified by an ID code followed by the motif length. The aligned motifs themselves are then provided, together with the corresponding source database ID code, the location in the parent sequence of each fragment (ST), and the interval between the fragment and its preceding neighbor (INT).

The protein domain families database, Pfam contains curated multiple sequence alignments for each family with linked functional annotation and profile hidden Markow models (HMMs) for finding the domains in the query sequences (Bateman *et al.*, 2000). A HMM is a probabilistic model that is suited for providing a mathematical scoring scheme for profile analysis. It describes the propensity of amino acid exchange in common protein domains and conserved regions. The database can be accessed interactively at http://www.sanger.ac.uk/Pfam/ by keyword search (enter keywords, e.g. enzyme name), Protein search (paste or upload query sequence) or DNA search (paste query sequence). Another HMM library is SMART, which can be accessed at http://smart.embl-heidelberg.de/. SMART is a web tool for the identification and annotation of protein domains and provides a platform for the comparative study of complex domain architectures in genes and proteins (Letunic *et al.*, 2006). The analytical results are returned with a graphical sketch of the predicted features and a summary list. PANAL (http://mgd.ahc.umn.edu/panal/run_panal.html) of Computational Biology Centers at the University of Minnesota, enables searches of several libraries to be performed simultaneously. These include ProSite, BLOCKS, PRINTS and Pfam (Bateman *et al.*, 2000).

16.2.4 Boutique databases

Many databases, known as boutique (specialized) databases select, annotate and recombine data focused on particular topics and include links affording streamlined access to information about subjects of interest, especially functional groups such as different families of enzymes (Table 16.4) and receptors (Table 16.5).

Another common group of boutique database is based on organisms. Yeast Proteome Database (YPD) maintained by Proteome Inc (http://www.proteome.com/) is currently the most comprehensive database for the budding yeast, *S. cerevisiae*. Proteome Inc has similar databases (DB) for *Candida albicans* and *Caenorhavditis elegans* with *Schizosaccharomyces pombe*, *Pneumocystis carinii, Aspergillus fumigtus* and *Aspergillus nidulans* in the development. Some of the other public organism DB include FlyBase (http://flybase.bio.indiana.edu/) and the *Arabidopsis thaliana* database, AatDB (http://genome-www.stanford.edu/

16.3 PROTEOME INFORMATICS: STRUCTURE DATABASES AND SERVERS

16.3.1 Structure database: Primary archive

Protein Data Bank (PDB) at http://www.rcsb.org/pdb/ is the worldwide archive of structural data of biomacromolecules (Kouranov *et al.*, 2006). PDB was established at

TABLE 16.4 Representative boutique enzyme databases

Database	URL
Aldehyde dehydrogenase	http://www.ucshc.edu/alcdbase/aldhcov.html
AARSDB: Aminoacyl-tRNA synthetases	http://rose.man.poznan.pl/aars/index.html
CAZy: Carbohydrate active enzymes	http://afmb.cnrs-mrs.fr/~pedro/CAZY/db.html
Esther: Esterases and hydrolases	http://www.ensam.inra.fr/esther
G6P dehydrogenase	http://www.nal.usda.gov/fnic/foodcomp/
HIV Proteases	http://www-fbsc.ncifcrf.gov/HIVdb/
KinG: S/T/Y-specific kinases	http://www.hodgkin.mbu.iisc.ernet.in/~king
Lipase Engineering DB: Lipases, esterases	http://www.led.uni-stuttgart.de/
LOX-DB: Lipoxygenases	http://www.dkfz-heidelberg.de/spec/lox-db/
MDB: Metalloenzymes	http://metallo.scripps.edu/
Merops: Peptidases	http://merops.sanger.ac.uk/
2-Oxoacid dehydrogenase complex	http://qcg.tran.wau.nl/local/pdhc.htm
PKR: Protein kinases	http://pkr.sdsc.edu/html/index.shtml
PlantsP: Plant protein kinases and phosphatases	http://PlantP.sdsc.edu
ProLysED	http://genome.ukm.my/prolyses/
Protease	http://delphi.phys.univ-tours.fr/Prolysis
REBASE: Restriction enzymes	http://rebase.neh.com/rebase/rebase.html
Ribonuclease P Database	http://www.mbio.ncsu.edu/RnaseP/home.html

TABLE 16.5 Representative boutique databases for receptor and signaling proteins

Database	URL	Description
Endo GPCR List	http://www.tumor-gene.org/GPCR/gpcr.html	GPCR, expression
GPCRDB	http://www.gpcr.org/7tm/	GPCR
gpDB	http://bioinformatics.biol.uoa.gr/gpDB	G-prot and interaction with GPCR
HORDE	http://bioinfo.weizmann.ac.il/HORDE/	Human olfactory R
LGICdb	http://www.pasteur.fr/recherche/banques/LGIC/LGIC.html	Ligand-gated ion channel subunit seq
NuclearRDB	http://www.receptors.org/NR/	Nuclear R superfamily
NR Resource	http://nrr.georgetown.edu/NRR/nrrhome.htm	Nuclear R superfamily
NUREBASE	http://www.ens-lyon.fr/LBMC/laudet/nurebase.html	Nuclear hormone R
PDB_TM	http://www.enzim.hu/PDB_TM/	Transmembrane prot with known 3D
Relibase	http://relibase.ebi.ac.uk/reli-cgi/rll?/reli-cgi/query/form_home.pl	R sequences, R binding sites
RTKdb	http://pbil.univ-lyon1.fr/RTKdb/	Receptor Y kinase seq
SENTRA	http://www.wit.mcs.anl.gov/sentra/	Sensory signal transduction prot
SEVENS	http://sevens.cbrc.jp/	7-Membrane helix R
VKCDB	http://vkcdb.biology.ualberta.ca/	Voltage-gated K-channel DB

Note: Abbreviations used: endo, endogenous; GPCR, G-protein-coupled receptor(s); prot, protein(s); R, receptor(s).

Brookhaven National Laboratories (BNL) in 1971 (Bernstein *et al.*, 1977). In October 1998, the management of the PDB became the responsibility of the Research Collaboratory for Structural Bioinformatics (RCSB) (Berman *et al.*, 2000; Deshpande *et al.*, 2005). Two methods, SearchLite (simple keyword search) and SearchFields (advanced search) are available to search and retrieve atomic coordinates (pdb files). If you know the PDB identification code (4 character identifier of the form, [0–9][A–Z or 0–9] [A–Z or 0–9]

[A–Z or 0–9] e.g. 1LYZ for hen's egg-white lysozyme and 153L for goose lysozyme), enter the identification code into the PDB ID box and hit the Explore button. The Structure Explore page for this entry is returned. If the PDB ID is not known, clicking Search-Lite opens the query page. Enter the keyword (name of ligand/biomacromolecule or author) and click the Search button. In the advanced search, clicking SearchField opens the query form. Construct combination of your query options including PDB ID, citation author, chain type (for protein, enzyme, carbohydrate, DNA or RNA), compound information, PDB header and experimental technique used. Clicking the Search button returns you to the search page with a list of hits from which you select the desired entry to access the Summary information of the selected molecule.

From the Summary information, the user can choose one of many options, including View Structure, Download/Display File, Structural Neighbors (links to CATH, CE, PSSP, SCOP and VAST), Geometry or Sequence Details. Select Download/Display File then choose PDB text and PDB noncompression format to retrieve the pdb file in text format. To view the structure online, select View Structure followed by choosing one of the 3D display options. The display can be saved as an image.jpg/image.gif. The Structure Neighbors provide links to CATH (fold classification by domain), CE (representative structure comparison, structure alignments, structure superposition tool), FSSP (fold tree, domain dictionary, sequence neighbors, structure superposition), SCOP (Class, fold, superfamily and family classification) and VAST (representative structure comparison, structure alignments, structure superposition tool). The Geometry option provides tables of dihedral angles, bond angles and bond lengths.

Molecular modeling database (MMDB) of Entrez is a subset of 3D structures from PDB recorded in asn.1 format (Wang *et al.*, 2002). The database that provides the link between 3D structures and sequences, can be accessed at http://www.ncbi.nlm.nih.gov/Entrez/structure.html. Enter author's last name or text words/keywords (e.g. zinc finger, DNA protein complex or topoisomerase) or PDB ID (if known) and click the Search or Go button. A list of hits is returned. Select the desired entries and then click the PDB ID to receive the MMDB Structure summary. Choose options (radios) to view or save the structure file as structure.cn3 (Cn3D), structure.kin (KineMage) or structure.pdb (RasMol). Sequence similarities or structure similarities can be viewed/saved by clicking the respective chain designation of Sequence neighbors or Structure neighbors. This returns the sequence display summary or VAST structure neighbors respectively.

WPDB (Shindyalov and Bourne, 1997), which can be downloaded from http://www.sdsc.edu/pb/wpdb, is a protein structural data resource. It is a Microsoft Windows-based, data management and query system that can be used locally to interrogate the 3D structures of biomacromolecules found in PDB. The program provides a query ability to find structures and performs sequence alignment, property profile analyses, secondary structure assignment, 3D rendering, geometric calculation and structure superposition (based on PDB atomic coordinates without computation). The accompanying WPDB loader (WPDBL) permits the user to build a subset of database from selected pdb files for use as tutorials on protein structures (Tsai, 2001).

The PDBsum (Laskowski, 2001; Laskowski *et al.*, 2005) is a Web-based compendium providing summaries and analyses of all structures in the PDB. Each summary gives an at-a-glance overview of the contents of a PDB entry in terms of resolution and R-factor, numbers of protein chains, ligands, secondary structure, fold cartoons and ligand interactions, etc. The facility draws together in a single resource information at the 1D (sequence), 2D (motif) and 3D (structure) levels. PDBsum is accessible for keyword interrogation at http://www.ebi.ac.uk/thornton-srv/databases/pdbsum. Search the structure by entering the four character PDB code (e.g. 1lyz, this returns the summary page) or a string of the protein name (e.g. lysozyme, this returns a list of hits. You have to select the desired entry to open the summary page). The summary page contains brief descriptions of the structure, hyperlinks to PDB, GRASS (Virtual graphic server at Colombia University), MMDB, CATH, SCOP, PROCHECK (Ramachandran plot statistics), PROMOTIF (summary of protein secondary structures at the site) and clickable presentations of CATH classification, secondary structures, PROMITIF, TOP, SAS (annotated FASTA alignment of related sequences in the PDB), PROSITE and LIGAND (ligand molecules and ligand binding residues). In addition to the graphical representation of secondary structures along the sequence, the structural information can be viewed by selecting/specifying type structural elements including helical wheels of PROMOTIF.

16.3.2 Structure databases: Substructures and structure classification

Many proteins share structural similarities due to the evolutionary process involving substitutions, insertions and deletions in amino acid sequences. Consequently protein structures can be characterized according to their common substructures (supersecondary structures, e.g. motifs, domains). For proteins with conserved functions, the structural environments of critical active site residues are also conserved. In an attempt to better understand sequence-structure relationships and the underlying evolutionary processes that give rise to different fold families, a variety of structure classification schemes have been established. Analyses of the 3D structures archived in PDB generate various databases for the specification/search of characteristic substructures and protein structure classifications (Table 16.6).

At iMolTalk (http://i.moltalk.org), the server offers:

1. general information such as header and sequence derived from 3D coordinates;
2. map corresponding residues from sequence to structure;
3. contact residues, heteroatoms of cofactors and ligands; and
4. identification of protein-protein interfaces.

SA-Search (http://bioserv.rpbs.jussieu.fr/cgi-bin/SA-Search) can be used to mine for protein structures and extract structural similarities. The structural alphabet (SA) allows the compression of 3D conformations into a 1D representation using a limited number of prototype conformations to permit the 3D similarity searches. STING Mellennium Suite (SMS) at http://trantor.bioc.columbia.edu/SMS provide analysis and visualization of sequence to structure relationships, such as nature and volume of atomic contacts, relative conservation of amino acids at the specific position and indication of folding essential residues, conformational map, rendering and coloring of the molecule.

The CATH (Class, Architecture, Topology, Homology) database is a hierarchical domain classification of protein structures (Pearl *et al.*, 2005) maintained at Biomolecular Structure and Modeling Unit of University College London, http://www.biochem.ucl.ac.uk/bsm/cath/. Different categories within the classification are identified by means of both unique numbers and descriptive names with five hierarchical levels. Class is derived from gross secondary structure content and packing. Four classes of domain are recognized:

1. mainly-α;
2. mainly-β;

TABLE 16.6 Some protein structure databases

Database	URL	Description
Primary archive		
PDB	http://www.rcbs.org/pdb/	Global 3D structure archive
Structure information		
iMolTalk	http://i.moltalk.org	Structural information
SA-Search	http://bioserv.rpbs.jussieu.fr/cgi-bin/SA-Search	Structure mining
SMS	http://trantor.bioc.columbia.edu/SMS	Sequence-structure relationships
Substructures		
ASTRAL	http://astral.stanford.edu/	Domains, seq-struct correspondence
LPFC	http://www-smi.stanford.edu/projects/helix/LPFC	Protein family core structures
PASS2	http://ncbs.res.in/~faculty/mini/compass/pass.html	Structural motifs
eF-site	http://ef-site.protein.osaka-u.ac.jp/eF-site	ESP, hydrophobic properties of AS
Het-PDB Navi	http://daisy.nagahama-i-bio.ac.jp/golab/hetpdbnavi.htmol	Heteroatoms in protein structures
Conformational analysis		
CADB	http://cluster.physics.iisc.emet.in/cadb/	Conformational angles
Decoys 'R' Us	http://dd.stanford.edu/	Comput generated prot conform
DisProt	http://divac.ist.temple.edu/disprot	Prot disorder without fixed 3D
DSDBASE	http://projects.villa-bosch.de/dbase/dsmm/	Simulated molecular motions
MolMovDB	http://bioinfo.mbb.yale.edu/MolMovDB/	Macromolecular movements
PepConfDB	http://www.peptidome.org/products/list.htm	Peptide conformations
Structure classification		
ArchDB	http://gurion.imim.es/archdb	Classif of loop structures
CATH	http://www/biochem.ucl.ac.uk/bsm/cath_new	Protein domain structures
CE	http://cl.sdsc.edu/ce.html	Nearest neighbors
FSSP	http://www.bioinfo.biocenter.helsinki.fi:8080/dali/	Protein fold classification
LPFC	http://www-smi.stanford.edu/projects/helix/LPFC	Protein family cores
MMDB	http://www.ncbi.nlm.nih.gov/structure	Neighbor analysis
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop	Structural classif of proteins
Sloop	http://www-cryst.bioc.cam.ac.uk/~sloop/	Classif of protein loops
SUPERFAMILY	http://supfam.org/	Structural superfamily

Note: Abbreviations used: AS, active site(s); classif, classification; comput, computer; conform, conformation(al); ESP, electrostatic potentials; prot, protein(s); seq, sequence(s); struct, structure(s).

3. α-β which includes both α/β and $\alpha + \beta$ structures; and
4. those with low secondary structure content.

Architecture describes the gross arrangement of secondary structures, ignoring their connectivities such as barrel, roll, sandwich, etc. Topology refers to both the overall shape and the connectivity of secondary structures. This is accomplished by means of structure comparison algorithms that use empirically derived parameters to cluster the domains. Homology groups domains that share equal or greater than 35% sequence identity and are considered to share a common ancestor. Similarities are first identified by sequence comparison and then by means of a structure comparison algorithm. Sequence clusters homology groups further on the basis of sequence identity such that domains have sequence identity greater than 35% indicating highly similar structures and functions.

The Structural Classification of Proteins (SCOP) database describes structural and evolutionary relationships between proteins of known structure (Lesk and Chothia, 1984;

Andreeva *et al.*, 2004). Proteins are classified in three levels hierarchically to refract their structural and evolutionary relatedness. Proteins are clustered into families with clear evolutionary relationships if they have sequence identities ≥30%. In the absence of significant sequence identity, in some cases, it is possible to infer common descent from similar structures and functions. Proteins are placed in superfamilies if their structural and functional characteristics suggest a common evolutionary origin in spite of low sequence identity. Protein are classed as having a common fold if they have the same major secondary structures in the same arrangement and with the same topology, whether or not they have a common evolutionary origin. SCOP is accessible for keyword interrogation via the MRC Laboratory of Molecular Biology Web server, http://scop.mrc-lmb.cam.ac.uk/scop/. The SCOP is augmented by the ASTRAL compendium (Brenner *et al.*, 2000) at http://astral.stanford.edu/ that provides sequence databases corresponding to the domain structures.

FSSP is the database that compares all PDB structures using the Dali program and processes them into domains. FSSP data can be accessed at http://www.bioinfo.biocenter.helsinki.fi:8080/dali/ by either searching for a PDB structure of interest or submitting the coordinates of a new structure and requesting DALI to perform a custom search against the FSSP database.

MMDB (http://www.ncbi.nlm.nih.gov/structure) is created at the National Centre for Biotechnology Information using a vector alignment search tool (VAST). Domains are first identified automatically on the basis of compactness. Similarities between domains are then determined by comparing sets of secondary structure vectors and assessing the significance of each hit with a statistical scoring scheme. By requesting structural neighbors, the user is effectively asking for structures whose scores with the target are the highest. The hierarchies for protein structure classification among these databases are:

CATH	SCOP	MMDB	FSSP
Class	Class		
Architecture			
Topology	Fold	VAST neighbor	Fold classification
Homologous superfamily	Superfamily		
Sequence family	Family	Sequence neighbor	Sequence homolog

16.4 PROTEOME INFORMATICS: PROTEOMIC SERVERS

16.4.1 Proteome analysis and annotation

Most proteomics tools available on web servers perform sequence analyses aimed at protein structure identifications/predictions and functional annotations. Some of these servers are given in Table 16.7.

The comprehensive proteomics server, ExPASy (Expert Protein Analysis System) at http://www.expasy.org is the main host of the SWISS-PROT knowledgebase (http://www.expasy.org/sprot/), SWISS-2DPAGE (http://www.expasy.org/ch2d/), PROSITE (http://www.expasy.org/prosite/), ENZYME (http://www.expasy.org/enzyme/) and SWISS-MODEL Repository (http://expasy.org/swissmod/smrep.html) maintained by the Swiss Institute of Bioinformatics (SIB) (Appel *et al.*, 1994; Gasteiger *et al.*, 2003). The server (Figure 16.3) provides Proteomic tools (http://www.expasy.org/tools) that can be accessed directly or from the ExPASy home page by selecting Identification and

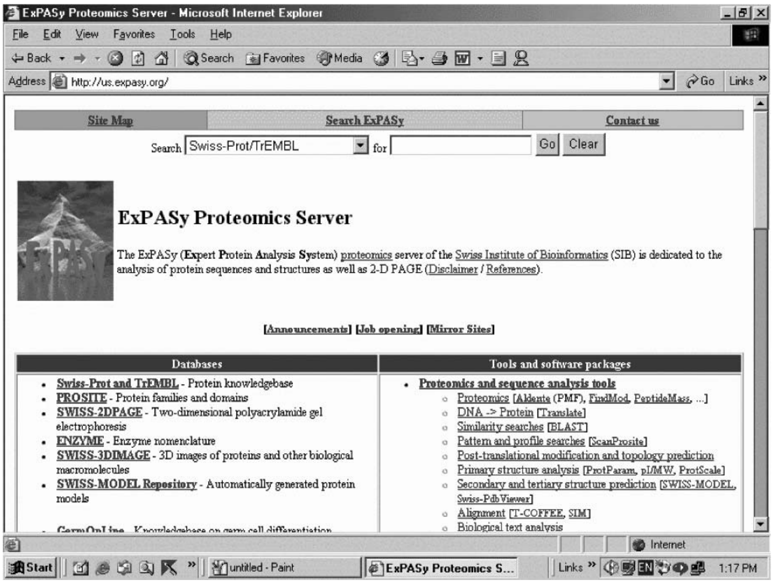
TABLE 16.7 Some proteomic web servers

Server	URL	Description
Proteomic analysis		
ExPASy tools	http://www.expasy.org/tools	Comprehensive proteomic tools
NPS@	http://npsa-pbil.ibcp.fr	Sec/cons structure prediction
CRASP	http://wwwmgs.bionet.nsc.ru/mgs/programs/crasp/	Substitution analysis in Malign
CHOP	http://www.rostlab.org/services/CHOP/	Structural domains
PipeAlign	http://igbmc.u-strasbg.fr/PipeAlign	Protein family analysis
PredictProtein	http://cubic.bioc.columbia.edu/predictprotein	Structure and function prediction
Motif3D	http://www.bioinf.man.ac.uk/dbbrowser/motif3d/motif3d.html	Sequence motifs
MotifViz	http://biowulf.bu.edu/MotifViz	Analysis of motifs
UniqueProt	http://cubic.bioc.columbia.edu/services/uniqueprot	Protein sequence sets
BioInf3D	http://bioinfo3d.cs.tau.ac.il	Structural bioinformatics
Proteome annotation		
Bioverse	http://bioverse.compbio.washington.edu	Funct and struct annotation of seq
InterPro	http://www.ebi.ac.uk/interpro	Proteome functional info
MISP	http://mips.gsf.de	Functional annotation
SMART	http://smart.embl.de/	Identification/annotation of domains
ProteomeAnalyst	http://www.ualberta,ca/~bioinfo/PA/	Protein properties and annotation
iSPOT	http://cbm.bio.uniroma2.it/ispot	Functional classification
ELM	http://elm.eu.org/	Eukaryotic functional sites
SIFT	http://blocks.fhcrc.org/shift/SIFT.html	AA change and protein function
SDPpred	http://math.genebee.msu.ru/~psn/algo.htm	Functionally specific AA
CD-Search	http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsd.cgi	Protein domain annotation, fly
MyHits	http://myhits.isb-sib.ch	Protein annotation, domain identif.
Membrane proteins/subcellular localization		
Membrane helix	http://cubic.bioc.columbia.edu/services/tmh_benchmark/	Membrane helix prediction
BPROMT	http://www.jenner.ac.uk/BPROMPT	Membrane protein prediction
ConPred II	http://bioinfo.si.hitosaki-u.ac.jp/ConPred2/	Transmembrane topology predict.
PRED-TMBB	http://bioinformatics.biol.uoa.gr/PRED-TMBB	β-Barrel outer membrane topology
MITOPRED	http://mitopred.sdsc.edu	Mitochondrial protein prediction
LOC3D	http://cubic.bioc.columbia.edu/db/LOC3d/	Subcellular localization
ESLpred	http://www.imtech.res.in/raghava/eslpred/	Eukaryote subcellular localization
TargetP	http://www.cbs.dtu.dk/services/TargetP/	Subcellular localization

Notes: 1. Abbreviations used: cons, consensus; euk, eukaryote(ic); GPCR, G-protein coupled receptor(s); sec, secondary; tert, tertiary; Malign, multiple alignments; phob, hydrophobic; prok, prokaryte(ic).
2. Mot structure prediction servers are given in Table 16.9.

Characterization under Tools and Software Packages. The Protein identification/characterization and sequence analysis tools of the Proteomics tools provide facilities to:

- identify a protein by its amino acid composition (AACompIdent);
- compare the amino acid composition of an entry with the database (AACompSim);
- compute theoretical p*I* and molecular weight (Compute pI/MW);
- predict potential post-translational modifications (FindMod);
- identify unexpected peptides (FindPept);
- predict potential protease/chemical cleavages (PeptideCutter);
- calculate the mass of peptides for peptide mapping (PeptideMass);


ExPASy Proteomics Server - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://us.expasy.org/
Go Links
Site Map
Search ExPASy
Contact us
Search Swiss-Prot/TrEMBL for
Go Clear
ExPASy
Internet
Start untitled - Paint ExPASy Proteomics S... Links 1:17 PM

ExPASy Proteomics Server

The ExPASy (**Ex**pert **P**rotein **A**nalysis **Sy**stem) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE (Disclaimer / References).

[Announcements] [Job opening] [Mirror Sites]

Databases

- **Swiss-Prot and TrEMBL** - Protein knowledgebase
- **PROSITE** - Protein families and domains
- **SWISS-2DPAGE** - Two-dimensional polyacrylamide gel electrophoresis
- **ENZYME** - Enzyme nomenclature
- **SWISS-3DIMAGE** - 3D images of proteins and other biological macromolecules
- **SWISS-MODEL Repository** - Automatically generated protein models

- **GermOnLine** - Knowledgebase on germ cell differentiation

Tools and software packages

- **Proteomics and sequence analysis tools**
  - Proteomics [Aldente (PMF), FindMod, PeptideMass, ...]
  - DNA -> Protein [Translate]
  - Similarity searches [BLAST]
  - Pattern and profile searches [ScanProsite]
  - Post-translational modification and topology prediction
  - Primary structure analysis [ProtParam, pI/MW, ProtScale]
  - Secondary and tertiary structure prediction [SWISS-MODEL, Swiss-PdbViewer]
  - Alignment [T-COFFEE, SIM]
  - Biological text analysis

Figure 16.3 Proteomic server page of ExPASy
The ExPASy Proteomic Server (http://www.expasy.org) or its mirror site (e.g. http://us.expasy.org in U.S.A. and http://ca.expay.org in Canada) provides comprehensive resources and tools for proteomic analyses. They include Databases (e.g. Swiss-Prot, Prosite, Enzyme), Tools (e.g. BLAST, ProtParam, ProtScale, SIM, SWISS-MODEL, Translate), Services (e.g. ExPASy ftp server, BioHunt, 2D Hunt), Documentation and links.

- identify protein using p*I*, MW, composition, sequence tags and peptide mass fingerprints (PeptIdent, TagIdent, MultiIdent);
- conduct sequence similarity searches (BLAST);
- evaluate physicochemical parameters from a query sequence (ProtParam);
- scan for PROSITE (ScanProsite);
- calculate/plot amino acid scales (ProtScale);
- generate random protein sequence from composition and length (RandSeq);
- predict tyrosine sulfation sites (Sulfinator), and
- translate nucleotide sequence into protein sequences in six reading frames (Translate).

In addition, two facilities related to glycomic tools, GlycanMass (calculate oligosaccharide mass) and GlycoMod (predict possible oligosaccharide structures on proteins) are also available. To use the Tools, for example AACompIdent, which performs the identification of a protein from its amino acid composition and physicochemical properties (e.g. pI and molecular weight), is a valuable step prior to initiating sequence determination. On the AACompIdent page, select a Constellation to be used (e.g. Constellation 0 for all amino acids) to open the request form. Enter pI value, molecular weight, biological source of the protein (OS or OC), the ExPASy recognizable keyword (e.g. dehydrogenase, kinase), your e-mail address and amino acid composition (in molar %). Click the Run AACompIdent button. The search result is returned by e-mail.

One of the popular server combining neural networks with multiple sequence alignments known as PHD (Rost and Sander, 1993) and a host of methods is available from

PredictProtein (Rost and Liu, 2004) at http://www.predictprotein.org. PredictProtein offers the comprehensive protein sequence analysis, structure and function predictions. These include database (e.g. SWISS-PROT + TrEMBL, PDB), alignment (e.g. BLAST, PSI-BLAST, HMMS, Prediction-based threading), domain identification (e.g. ProDom, Pfam, CHOP, etc.), structure prediction (e.g. PHDs, GLOBE for globularity, COILS for coiled-coil regions, etc) and functional assignment (e.g. PROSITE, LOCnet for subcellular localization, etc). From the submitted protein sequences, PredictProtein returns multiple sequence alignments, PROSITE sequence motifs, low-complexity regions, nuclear localization signals, regions lacking regular structure, predictions of secondary structures, solvent accessibility, globular regions, transmembrane helices, coiled-coil regions, structural switch regions, disulfide bonds, subcellular localization and functional annotations. Upon request, fold recognition, domain assignments, predictions of transmembrane strands and inter-contacts residues are also available.

InterPro (http://www.ebi.ac.uk/interpro) is an integrated documentation resource of protein families, domains and functional sites by combining major signature databases, PROSITE, Pfam, PRINTS, SMART, TIGRFAMs, PIRSF and SUPERFAMILY. The resource provides biological, functional and taxonomic information, gene ontology mapping, protein match views and protein 3D structure data.

Protein sequences are structurally and functionally annotated at Bioverse (http://bioverse.compbio.washington.edu/). The annotations consist of three sections (McDermott and Samudrala, 2003). The sequence section lists similar sequences identified by searches using various methods, including PSI-BLAST. The structure section is composed of secondary and tertiary structure information. The function section combines PROSITE, BLOCK, PRINT, Pfam, ProDOM, SMART and TIGRFAM to match sequences to patterns, domains and families.

TargetP (http://www.cbs.dtu.dk/services/TargetP/) assigns the subcellular location of proteins based on the predictions of the N-terminal chloroplast transit peptide, mitochondrial targeting peptide or secretory signal peptide.

16.4.2 Integrated databases

The proteome research takes a big stride forward in the post-genome era. Functional genomics involve studies of transcript structures and expression by high-throughput transcriptomic techniques such as expression microarrays. However, transcripts are not directly operating molecules but are translated into functional proteins. Proteins need to be analyzed by proteomic techniques. Therefore functional genomics studies generally involve proteomics. Efforts have been made to integrate the genome information with the proteome information. The number of such databases that integrate informatics of genomes and proteomes from various organisms will no doubt increase with time, some of which are listed in Table 16.8.

The Integr8 portal (http://ebi.ac.uk/integr8) offers an overview of information about organism with completely sequenced genomes, statistical analysis of their genomes/proteomes individually and comparatively. The database is build on three main sources (Kersey *et al.*, 2005):

1. genome reviews of annoted genome sequences from multiple sources;
2. nonredundant sets of UniProt entries representing completed proteomes; and
3. IPI (International Protein Index) with comprehensive protein sets for certain higher metazoan species.

TABLE 16.8 Integrated genome/proteome databases

Database	URL	Integrated resources
Integr8	http://ebi.ac.uk/integr8	Genome/nucleotide sequences and annotations (EBI), gene expression (CleanEx), promoters (EPD), splice sites (RealSplice), transcription factors (TRANSFAC), UTRs (UTRdb), Protein sequences and annotations (UniProt), protein clusters (CluSTr), domains, families and repeats (InterPro), protein structures (PDB), gene ontology (GO) and annotations (GOA).
InterPro	http://www.ebi.ac.uk/interpro	Integration of all the domain families from established databases such as Pfam, PRINTS, SMART, Tigrfam, SCOP and CATH and mapping them onto the genome sequences.
NCBI: RefSeq	http://www.ncbi.nlm.nih.gov/RefSeq	Curated NR sequences: genomes, transcripts and proteins derived from GenBank, EBI, DDBJ and model organism databases (e.g. TIGR, MGI, SGD, FlyBase) with wide taxonomic diversity.
PBIL	http://pbil.univ-lyon1.fr	Nucleotide sequences (GenBank, EMBL), protein sequences (PIR, Swiss-Prot + TrEMBL), homologous genes (HOVERGEN, HOBACGEN), receptors (NUREBASE, RTKdb), genomes (EMGLib, NRSub) and PDB entries (NPSA_3Dseq).

Note: References cited are Kersey *et al*. (2005) for integr8; Pruitt, K.D. *et al*. (2005) for NCBI RefSeq; Perrière *et al*. (2003) for PBIL.

RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq) can be accessed via NCBI Entrez providing nonredundant sequences representing genomic data, transcripts and proteins of significant taxonomic diversity (Pruitt *et al.*, 2005). Explicit cross-links between nucleotide and protein cognates for most organisms are included. The genomic records are annotated with genes, transcripts and proteins with additional features. Transcript records (subset of eukaryotic species) include protein-coding sequences, transcribed pseudogenes, ribosomal RNA and other small RNAs. The goal of RefSeq is to represent the full-length protein product of the genome although partial protein products are also represented. Protein annotation includes automatic calculation of the conserved domains and other landmark regions of the protein sequences and references. The PBIL (http://pbil.univ-lyon1.fr) provides online access to nucleic acid/protein sequence databases and many analytical tools (Perriére *et al.*, 2003). The server allows users to query nucleotide sequences in the EMBL and GenBank formats and protein sequences in the SWISS-PROT and PIR formats. Some of the analysis tools include similarity search, pairwise/multiple alignments, phylogenetic tree visualization, protein structure prediction (e.g. Geno3D for molecular modeling and NPS@ via NPSA link page for secondary structure predictions), and multivariate statistics such as correspondence analysis (CA), principle component analysis (PCA) or discriminant correspondent analysis (DCA).

16.4.3 Post-translational modifications and functional sites

An important aspect in the sequence analysis of proteins is the detection of functional sites such as the active site, ligand binding site, signal sequence/cleavage site and post-translational modification sites. Generally, a prior knowledge concerning the chemical nature of these sites is available. For example, the active triad (Ser-His-Asp/Glu) of serine proteases, the Arg residue of NADP$^+$ binding site and the Asn/Gln residue of glycosylation site are known and detection of these functional sites can be accomplished by the alignment of a query sequence against sequences with known candidate sites. When studying the specificity of molecular functional sites, it has been common practice to create consensus sequences from alignments or structure comparison, and then to choose the most common amino acid residue(s) representative at the given position(s)/structures. Such knowledge is essential for the functional annotation of proteomes. The known functional sites are described in the FEATURE lines of the curated sequence files, such as GenPept, PIR and SWISS-PROT. Many online databases and servers are dedicated to the search and identification of functional, as well as posttranslational modification (PTM) sites (Table 16.9).

The PROSITE database (http://www.expasy.org/sprot/prosite.html) consists of biologically meaningful signatures that are described as patterns and profiles. Each signature is linked to documentation that provides useful biological information on the protein family, domain or functional site (Hulo *et al.*, 2004). The database that is composed of two asci (text) files, PROSITE.dat and PROSITE.doc, and can be downloaded using FTP from ExPASy (ftp://www.expasy.ch/databases/prostie/) or Swiss Institute for Experimental Cancer Research (ISREC) at ftp://ftp.isrec.isb_sib.ch/sib-isrec/profiles or EBI (ftp://ftp.ebi.ac.uk/pub/databases/profiles/). The similarity search for the highly conserved sequences/regions, which serve to identify the functional/signature sites, is available from the BLOCKS database (Henikoff *et al.*, 1998) at http://www.blocks.fhcrc.org/. Pfam at http://www.sanger.ac.uk/Pfam is a protein family database containing functional annotation (Bateman *et al.*, 2000). The on-line applications of these servers have been described.

The Catalytic Site Atlas (CSA) at http://www.ebi.ac.uk/thornton-srv/databases/CSA provides catalytic residue annotation for enzymes in the PDB (Porter *et al.*, 2004). The residues are defined as catalytic if they fulfill any one of the criteria; namely:

1. direct involvement in the catalytic mechanism;
2. alteration of the pKa of a residue or water molecule directly involved in the catalytic mechanism;

TABLE 16.9 Online resources for functional and post-translational modification sites

Web site	URL	Description
CBS	http://www.cbs.dtu.dk/	PTM, signal peptides
CSA	http://www.ebi.ac.uk/thornton-srv/databases/CSA	Catalytic residues in enzymes
Enz Structure	http://www.biochem.ucl.ac.uk/bsm/enzymes/index.html	Enz structure topology, active sites
O-GlycBase	http://www.cbs.dtu.dk/databases/OGLYCBASE/	O- & C-linked glycosylation sites
PDBSite	http://srs6.bionet.nsc.ru/srs6/	3D Structure of functl sites
PDBSiteScan	http://wwwmgs.bionet.nsc.ru/mgs/systems/fastprot/pdbsitescan.html	Active/binding/PTM site search
PEST	http://www.icnet.uk/LRITu/projects/pest/	PEST sequences
Phospho.ELM	http://phospho.elm.eu.org/	S/T/Y phosphorylation sites
dbPTM	http://dbPTM.mbc.nctu.edu.tw/	Experimentally validated PTM sites
PRECISE	http://precise.bu.edu/precisedb/	Consensus interact sites in enzymes
PROSITE	http://ww.expasy.org/prosite	Funct/signature patterns and profiles
PROMISE	http://metallo.scripps.edu/PROMISE	Prosthetic, metal binding sites
WebFEATURE	http://feature.standard.edu/webfeature	3D physiochemical motifs

Notes: 1. The table lists Web sites, both databases and servers that provide sequence/structure analyses to search/identify functional and PTM sites.

2. Abbreviations used: PTM, post-translational modification(s); S/T/Y, serine/threonine/tyrosine.

3. stabilization of a transition state or intermediate by lowering the activation energy for a reaction; and
4. activation of the substrate in some way.

PRECISE (http://precise.bu.edu/) identifies the consensus interaction sites for various ligands in enzymes that include binding of substrates, products, cofactors and inhibitors (Sheu *et al.*, 2005).

The active site topology of enzymes can be displayed/saved by accessing the Enzyme Structure Database (http://www.biochem.ucl.ac.uk/bsm/enzymes/index.html) and the binding site of receptors can be analyzed at Relibase (http://relibase.ebi.ac.uk/reli-cgi/rll?/reli-cgi/query/form_home.pl), as discussed earlier. Center for Biological Sequence Analysis (CBS) at http://www.cbs.dtu.dk/ offers methods for predicting posttranslational modifications including signal peptide cleavage (Nielsen *et al.*, 1999), O-glycosylation (Gupta *et al.*, 1999) and phosphorylation (Kreegipuu *et al.*, 1999) sites. OGlycBase returns tables of potential versus threshold assignments for Ser and Thr residues as well as a plot of *O*-glycosylation potential versus sequence position. Phospho.ELM returns tables of context (nanopeptides, [S,T,Y] ± 4 residues) and scores for Ser/Thr/Tyr predictions.

Proteins with an intracellular half-life ($t_{1/2} = 0.693/k$, where k is the first-order rate constant for the proteolytic degradation of a cellular protein) of less than two hours, are found to contain regions rich in proline, glutamic acid, serine and threonine termed PEST (Rodgers *et al.*, 1986). The PEST region is generally flanked by clusters of positively charged amino acids. The PEST sequence search is available at http://www.icnet.uk/LRITu/projects/pest/. Candidate PEST sequences with hydrophobicity index and PEST score are returned.

16.5 PROTEIN STRUCTURE ANALYSIS USING BIOINFORMATICS

There is a considerable impetus to predict accurately protein structures from sequence information because of the protein sequence/structure deficit as a consequence of the genome and full-length cDNA sequencing projects. The molecular mechanical (MM) approach to modeling of protein structures has been discussed in section 9.2, and the protein secondary structure prediction from sequence by statistical methods has been treated in section 9.5. The prediction of protein structure using bioinformatic resources will be described in this subsection. The approaches to protein structure predictions from amino acid sequences (Tsigelny, 2002; Webster, 2000) include:

- Secondary structure prediction without attempting to assemble these regions in three-dimensions. The results are lists of regions of the sequence predicted to form α-helices, β-strands and coils. In some cases, the propensities to bury/exposure are also reported.
- Fold recognition by determining which regions of a query protein of known sequence (but unknown structure) share a folding pattern of protein structure(s) in the library. Two approaches, structure superposition (overlap) and threading are generally employed. The results, if found, are a nomination of a known structure that has the same fold as the query protein.
- Homology modeling is performed to produce model(s) for the query protein based on the known 3D structure(s) of one or more related proteins. The results are a complete coordinate set for main-chain and side-chains, intended to be a high quality model of the 3D structure.

- Prediction of the 3D structure model, either by a prior or knowledge-based methods. The results are complete coordinate set for the main-chain and sometimes the side-chains.
- Refinement of the predicted model. Computational molecular modeling including energy minimization and dynamic simulation is carried to refine the model(s) obtained from the comparative structure modeling.

16.5.1 Secondary structure predictions

An approach to secondary structure predictions has been discussed (subsection 9.5.2). A number of Web sites (Table 16.10) offer services in the secondary structure predictions of proteins using different approaches with varied accuracy. The Secondary structure prediction of ExPASy Proteomic tools (http://www.expasy.ch/tools/) provides pointers to different Web servers for predicting secondary structures of proteins. The ProtScale of ExPASy Proteomic tools produces conformational profiles by plotting statistical scales of various parameters (e.g. Chou and Fasman's conformational propensities; Levitt's conformational parameters) against residue positions.

Numerous prediction methods including SOPM, SOPAMA, HNN, MLRC, DPM, DSC, GOR I, GOR III, GOR IV, PHD, PREDATOR and SIMPA96 are available at Network Protein Sequence Analysis (NPS@) (Combet *et al.*, 2000). The server can be accessed via http://npsa-pbil.ibcp.fr. For example, the prediction method of Garnier, Osguthorpe and Robson (GOR I, III or IV) can be initiated by selecting GOR I, III or IV under Secondary structure prediction tool to open the query form. Paste the query sequence into the sequence box and click the Submit button. The result is returned showing the query sequence with the corresponding predicted conformations (c, h, e, and t for coil, helix, extended strand and turn respectively), summarized content (%) of conformations and conformational profiles. The user may request the secondary structure consensus prediction enabling the simultaneous execution of a number of selected prediction methods. The predicted secondary structures including the consensus secondary structure (Sec.Cons.) are returned.

The Baylor College of Medicine (BCM) at http://www.hgsc.bcm.tmc.edu/Search-Launcher offers segment-oriented prediction (PSSP/SSP), in which the most probable secondary structure segments (a for α-helix, b for β-strand and c for the remainder) are assigned based on the probability for a, b or c (0 to 9). The PredSS assignments of a and

TABLE 16.10 Online servers for predicting protein secondary structures

Server	URL	Description
BCM	http://www.hgsc.bcm.tmc.edu/search_launcher/	Prediction
DSSPcont	http://cubic.bioc.columbia.edu/services/DSSPcont	Assignment
ExPASy	http://www.expasy.org/tools	Tools/links
LIBRA I	http://www.ddbj.nig.ac.jp/E-mail/libra/LIBRA_I.html	Secs and accessibility
nnPredict	http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html	Identification/class
NPS@	http://npsa-pbil.ibcp.fr	Diff pred methods
Predator	http://www.embl-heidelberg.de/cgi/predator_serv.pl	Secs and accessibility
PSA	http://bmerc-www.bu.edu/psa/index.html	Secs and accessibility
PSIpred	http://insulin.brunel.ac.uk/psipred/	Secs and accessibility
SSThread	http://www.ddbj.nig.ac.jp/E-mail/ssthread/www_service.html	Sequence threading
STRIDE	http://bioweb.pasteur.fr/seqanal/interfaces/stride-simple.html	Secs assignment

Note: Abbreviations used: diff, Different; pred, prediction; secs , secondary structure.

b are made. The nnPredict (http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html) uses the neural network in the structure prediction and reports both the predicted secondary structures (H = helix, E = strand and – = no prediction) and tertiary structure class (all-α, all-β, and α/β). The Protein Sequence Analysis (PSA) server of BMERC predicts secondary structures and folding classes from a query sequence. On the PSA home page at http://bmerc-www.bu.edu/psa/index.html, select Submit a sequence analysis request to submit the query sequence and your e-mail address. The returned results include:

- probability distribution plots (conventional X/Y and contour plots) for strand, turn and helix; and
- a list of structure probabilities for loop, helix, turn and strand for every amino acid residues.

The prediction of the secondary structures can be made by the structure similarity search of PDB collection at the site. Several servers provide such prediction method. The Jpred which aligns the query sequence against PDB library, can be accessed at http://jura.ebi.ac.uk:8888/index.html. To predict the secondary structures, however, check Bypass the current Brookhaven Protein Database box and then click Run Secondary Structure Prediction on the home page of Jpred to open the query page. Upload the sequence file via browser or paste the query sequence into the sequence box. Enter your e-mail address (optional) and click the Run Secondary Structure Prediction button. The results with the consensus structures are returned either online (linked file) or via e-mail (if e-mail address is entered).

The 3D-1D compatibility algorithm (Ito *et al.*, 1997) is applied to predict the secondary structures by threading at SSThread of DDBJ (http://www.ddbj.nig.ac.jp/E-mail/ssthread/www_service.html). The threading result reports the amino acid sequence with the predicted secondary structures (H for α-helix; E for β-strand and C for coil or other). The threading prediction of the query sequence against the structural library chosen from PDB at LIBRA I (http://www.ddbj.nig.ac.jp/E-mail/libra/LIBRA_I.html) reports a list of compatible structures and 3D-1D alignment results recording the secondary structures and accessibility.

16.5.2 Three-dimensional structure modeling

Three categories of methods are applied to model 3D protein structures, namely comparative (homology) modeling (CM), fold recognition (FR) and *ab initio* (*de novo*) prediction (AB). In the CM and FR categories, the methodologies rely on the presence of sequence homologous proteins of known structures, which are used as the template(s) to construct a model. This can be accomplished by sequence similarity (Greer, 1990; Murzin and Bateman, 1997) or by threading a sequence against a library of structures and selecting the best match (Bovie *et al.*, 1991; Jones *et al.*, 1997). Sequence alignment between the target protein and the template protein(s) of known structure(s) is used to create the initial model, which is subsequently refined/validated. Various steps taken to construct 3D models of protein structures to atomic detail are shown in Figure 16.4. The Protein Model Database (PMDB) at http://www.caspur.it/PMDB offers an access to published and validated 3D models of proteins (Castrigano *et al.*, 2006).

In the *de novo* category, the prediction methods are based on general principles that govern protein structure and energies (Lee *et al.*, 1999; Ortiz *et al.*, 1999; Samudrala *et al.*, 1999). The categories vary in difficulty and levels of sophistication. Consequently various methods generate 3D models with different degrees of accuracy relative to the

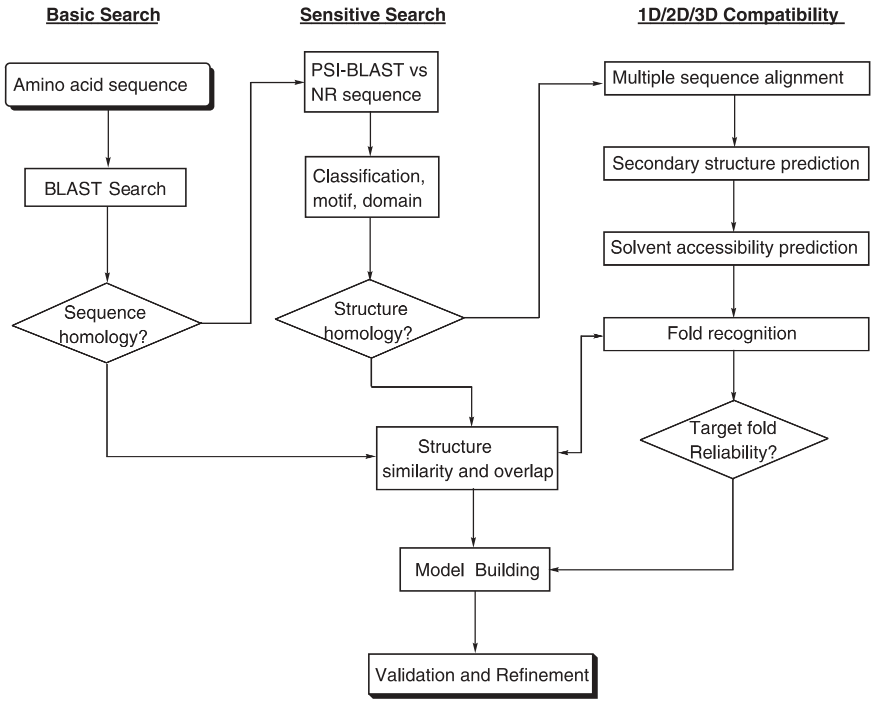

Basic Search
Sensitive Search
1D/2D/3D Compatibility
Amino acid sequence
BLAST Search
Sequence homology?
PSI-BLAST vs NR sequence
Classification, motif, domain
Structure homology?
Multiple sequence alignment
Secondary structure prediction
Solvent accessibility prediction
Fold recognition
Target fold Reliability?
Structure similarity and overlap
Model Building
Validation and Refinement

Figure 16.4 Protein structure prediction flowchart
Basic steps involved in protein structure prediction from its sequence by using bioinformatics are highlighted. The first step in a typical protein structure prediction is to establish if the query sequence (target sequence) has any structure homolgs in the Protein Data Bank (PDB) using sequence similarity searches/recognition. Comparative molecular modeling is carried out based on the homologous structure (template) to construct 3D models.

experimental structures. Structure prediction methods are systematically assessed through blind trials in the Critical Assessment of Techniques for Protein Structure Prediction (CASP). In general CM and FR can produce backbone structure with an accuracy of 4Å root-mean-square (RMS) distance as demonstrated in the CASP predictions (Moult *et al.*, 2003).

16.5.3 Sequence similarity and alignment

The structural similarity search can be conducted at any of the primary sequence databases that provide BLAST and PSI-BLAST searches. The search can be performed also at dedicated search servers such as Jpred (http://jura.ebi.ac.uk:8888/index.html) or G to P server (http://spock.genes.nig.ac.jp/~genone/gtop.html). For example, the G to P server conducts a search of a query sequence against the sequences of 3D structures (BLAST against PDB or PSI-BLAST against SCOP) by returning a search summary with multiple alignment.

Each amino acid has distinct attributes such as size, hydrophobicity/hydrophilicity, and hydrogen bonding capacity and conformational preferences that allow it to contribute to a protein fold. Attempts are being made to interpret the protein fold in terms of amino acid descriptors, and the local physicochemical environment of every residue within each 3D structure is directly obtainable (e.g. AAindex at http://www.genome.ad.jp/aaindex/ or

ProtScale of ExPASy Proteomic tools at http://www.expasy.ch/tools/). In sequence alignments, a gap penalty is typically assigned as either a constant value or one that consists of two components, a constant component and a length-dependent component. However, gaps do not occur with equal frequency everywhere within the protein fold. They are rarely found within elements of regular secondary structure, but rather most often within loop regions that are exposed to solvent at the surface of proteins. Likewise, a hydrogen-bonded polar amino acids buried in the protein core is highly conserved, whereas these same residues at the surface of a protein that is exposed to solvent, are not.

It is generally accepted that there are a finite number of different protein folds and it is likely that a newly determined structure will adopt a familiar fold rather than a novel one. This limitation to protein folds presumably derives from the relatively few secondary structure elements in a given domain and the fact that the arrangement of these elements is greatly constrained by folding environments, protein functions and probably evolution. Because proteins sharing at least 30% of their amino acid sequence will adopt the same fold and will often exhibit similar functions, protein homology has been an important quality in the identification of related proteins with common folds during database searches. Thus a sequence can be compared with the known 3D structure of homologous protein using one of two basic strategies (Johnson *et al.*, 1996). In the first approach, one-dimensional sequences are compared with one-dimensional templates or profiles derived from known 3D structures. The template provides a set of scores for matching a position within the structure with those in the sequence. Thus threading methods (Jones *et al.*, 1992) fit a probe sequence on to a backbone of a known structure and evaluate the compatibility by using pseudo-energy potential and residue environmental information (Shi *et al.*, 2001). The second strategy involves an evaluation of the total energy of a sequence of the molecule, which is folded similarly to the known 3D structure.

When representative structures for most of the common folds become available, the structure alignment will certainly be the most powerful approach to protein fold recognition. Thus computer graphics, with its interactive capabilities, emerged as an attractive tool for displaying molecular shapes or superimposing active molecules to detect their common substructures. For two related structures, the main question is finding the best geometric transformation in order to superimpose their frameworks together as closely as possible and evaluate their degree of similarity. If a protein with a known sequence but unknown structure can be matched to one of the common folds, it is possible to model the 3D structure of the protein and learn something about its function.

16.5.4 Structure similarity and overlap

With protein structures, we can observe similarities in topology or even in structural details. Comparison of protein structures can reveal distant evolutionary relationships that would not be detected by sequence alignment alone. Comparing and overlapping two protein structures quantitatively remain an active area of development in structural biochemistry. Methods for protein comparison generally rely on a fast full search of the protein structure database.

In devising measures of similarity and difference between two proteins, it is sometimes clearer to note how to proceed in comparing sequences, than to do so in comparing structures. If the amino acid sequences of two proteins can be aligned, then we can either count the number of identical residues, or use a similarity index between amino acids. Such an index would take the form of a 20 × 20 matrix, M, such that each entry corresponds to a pair of amino acids and M_{ij} gives a measure of the similarity between any pair of amino acids. The similarity between two sequences is then the sum of values for each pair of aligned amino acids taken from the matrix, plus a correction to account for the gaps found in the sequences at sites of insertions or deletions of amino acids (Needleman and Wunsch, 1970). Indeed, it is by maximizing such a similarity score that the optimal alignment of two sequences is conventionally calculated.

In three dimensions the problem can be more complex. If two protein structures are closely related, and we align the residues, it is then easy to superpose the corresponding residues, and to measure and analyze the nature of the deviations in position of corresponding atoms. Structure tends to change more conservatively than sequence, and it is not uncommon to be able to recognize a relationship between proteins from their structures, when no evidence of homology appears in the sequences (Brändén, 1980; Levitt and Chothia, 1976). Using computer graphics, it is possible to superpose two (or more) structures in one picture, and this can reveal immediately which features are conserved and which are different between two related proteins or in the conformational changes during allosteric transition/upon complexation. It is this facility in which lies the real advantage of computer graphics.

There are two ways to superpose structures. We can display pictures of each structure and use interactive graphics to provide the facility to rotate and translate one with respect to the other, superposing the two ‘by eye’. The second method is numerical. By selecting corresponding sets of atoms from two structures, a program can calculate the best ‘least square fit’ of one set of atoms to another. Both methods are useful: Fitting by eye permits rapid experimentation to assess the goodness of fit of different portions of the molecules. Numerical fitting permits quantitative comparisons of the relative goodness of fit of different structures and substructures.

Suppose we are dealing with two or more protein structures that contain regions in which the backbone atoms are almost congruent. We wish to superpose the two structures by moving one with respect to the other so that the corresponding atoms in the well-fitting regions are optimally matched. There are two aspects to this problem:

1. choosing the corresponding sets of atoms/substructures from the two structures; and
2. finding the best match of the corresponding atoms/substructures.

The algorithms fall into two broad types:

1. those that attempt to optimally superpose structures by minimizing the intermolecular distances between equivalent positions; and
2. those that compare intramolecular distances between residues or secondary structures in two proteins.

Various optimization strategies are employed to align the intermolecular or intramolecular distances, including simulated annealing, Monte Carlo optimization, and dynamic programming (Brown *et al.*, 1996). Although most methods give similar alignments for closely related proteins, alignments can be sensitive to the features used to identify equivalent positions and considerable differences can be observed for distant homologues as their structures diverge. Generally, methods agree in their identification of equivalent secondary structures. Loop regions are the most difficult to align. The 3D-template based approach attempts to capture the most conserved structural characteristics of protein folds by compiling a library of structural cores to facilitate the recognition of structural similarity.

Rigid-body approaches are used for measuring the root-mean-square deviation (RMSD) or average distance between two protein structures after superposition of equivalent positions (Rossmann and Argos, 1975). Alternatively, the trimmed RMSD uses only equivalent pairs (core elements) within a threshold distance from each other to calculate RMSD (May, 1999). The basic idea is that if some set of corresponding residues from each of two proteins fits well, then any subsets of corresponding residues will also fit well, though the converse is not necessarily true. For two related structures, the main question is finding the best geometric transformation in order to superimpose their frameworks together as closely as possible and evaluate their degree of similarity. Generally, the superimposition is carried out by finding the rigid body translation and minimizing the root-mean-square (RMS) distance between corresponding atoms of the two molecules, i.e. two structures are superimposed and the square root is calculated from the sum of the squares of the distances between corresponding atoms:

$$\mathrm{RMS(rms)} = \left\{ \sum_{i}^{N} (|u_i - v_i|)^2 \right\}^{1/2}$$

where u_i and v_i are corresponding vector distances of the ith atom in the two structures containing N atoms. The result is a measure of how each atom in the structure deviates from each other and an RMS value of 0–3 Å signifies strong structural similarity.

It is useful to set a threshold for goodness of fit, for example 0.5 Å. By recording only those pairs of substructures for which RMS deviation, $\Delta\mathrm{RMS} \leq 0.5$, we has generated a list of pairs of well-fitting substructures. The next step is to work toward larger substructures by merging entries in this list. The procedure can be iterative, given a list of pairs of well-fitting substructures, form the unions of pairs of entries, fit them and then append to the list any new well-fitting substructures discovered. Another approach is to characterize the conformation by the sequence of main-chain conformational angles ϕ and ψ and extract sets of consecutive residues with similar conformations (Levine *et al.*, 1984). To consider the extent or range of the structural similarity, RMSD or any appropriate scores (e.g. raw alignment score, position-weighted superposition score or Z-score) are usually quoted together with the number of equivalent residues identified. Some of these methods available online are listed in Table 16.11.

TABLE 16.11 Online servers for protein structure alignment, similarity/overlap

Servers	URL	Description
BioInfo3D	http://bioinfo3d.ca.tau.ac.il	Alignment, compar, interactions
CE	http://cl.sdsc.edu/ce.html	Comparison and alignment
CE-MC	http://cemc.sdsc.edu	Multiple structure alignment
CKAAPs DB	http://ckaap.adsc.edu/	Similar struct with seq dissimilar
FATCAT	http://fatcat.ljcrf.edu	Similarity search/comparison
FoldMiner	http://foldminer.stanford.edu/	Similarity search and alignment
HOMSTRAD	http://www-cryst.bioc.cam.ac.uk/homstrad	Homologous struct alignments
LGA	http://PredictionCenter.llnl.gov/local/lga	Structure comparison
MATRAS	http://biunit.aist-nara.ac.jp/matras	3D Structure comparison
ProteinDBS	http://ProteinDBS.rnet.missouri.edu	Structure comparison
SS DB	http://ssd.rbvi.ucsf.edu/	Superposition of barrel structures
SuperPose	http://wishart.biology.ualberta.ca/SuperPose	Structural superposition
Wurst	http://www.zbh.uni-hamburg.de/wurst	Sequence to structure alignment

16.5.5 Fold recognition and threading

Families of related proteins tend to retain similar folding patterns. If we examines sets of related proteins, it is clear that general folding pattern of the structural core is preserved. Proteins having significant sequence similarity (above 30% identity) adopt similar folds (Chothia and Lesk, 1986). Protein family analyses have suggested that 50–60% of the protein structure is conserved amongst relatives, generally in the core (Koppensteiner *et al.*, 2000). Certain motifs recur in many folds within a given class (e.g. α-hairpin in all-α β-hairpin, β-meander and Greek keys in mainly-β and βαβ motif in α/β) and may be associated with favorable physicochemical constraints on secondary structure packing that restrict their arrangements in 3D space. But there are distortions that increase in magnitude as the amino acid sequences diverge. In proteins, the common core generally contains the major elements of secondary structures and segments flanking them, including active site peptides. Large structural changes in the regions outside the core make it difficult to measure structural change quantitatively by straightforward application of simple least-squares superposition techniques. To define a useful measure of structural divergence, it is necessary first to extract the core and then carry out the least-squares superposition on the one alone (Chothia and Lesk, 1986). For each major element of secondary structure of two related proteins, a succession of superposition calculations, which include the main-chain atoms (N, Cα, C, O) of corresponding secondary structural elements plus additional residues extending from either end, is conducted. More and more additional residues, now identified as a well-fitting contiguous region containing an element of secondary structure plus flanking segments, are then included. After finding such pieces corresponding to all common major elements of secondary structure, a joint superposition of the main chain of all of them is performed. The structure superposition ensures an efficient fold recognition.

Threading is a method for fold recognition. The query protein of a known sequence but unknown structure is compared with a fold library, which can be some or all of PDB or even hypothetical folds. In threading, many rough models of the query protein are built based on each of the known structures using different possible alignments of the sequences. This systematic exploration of the many alignments produces a number of possible fold models (e.g. superfolds). InDels are allowed in the alignments. Both threading and homology deal with the 3D structure derived from the alignment of the query sequence with known structures of homologous proteins. Threading explores many alignments and deals with many and rough models, usually without constructing an explicit 3D models, whereas homology modeling focuses on one set of alignments and then aims at constructing a detailed 3D model. A number of threading Web servers are available, such as 3D-PSSM at http://www.bmm.icnet.uk/~3dpssm/ (Kelley *et al.*, 2000) and FUGUE at http://www-cryst.bioc.cam.ac.uk/~fugue/prfsearch.html (Shi *et al.*, 2001).

16.5.6 Homology modeling

One of the ultimate goals of protein modeling is the prediction of 3D structures of proteins from their amino acid sequences. The prediction of protein structures rely on two approaches that are complementary and can be used in conjunction with each other:

1. Knowledge-based model combining sequence data to other information, such as homology modeling (Hilbert *et al.*, 1993; Chinea *et al.*, 1995).
2. Energy-based calculations through theoretical models and energy minimization, such as *ab initio* prediction (Bonneau and Baker, 2001).

The most promising methodology relies on modification of a closely related (homologous sequence), functionally analogous molecule whose 3D structure has been elucidated. This is the basis of homology modeling for deriving a putative 3D structure of a protein from a known 3D structure. Residues are changed in the sequence with minimal disturbance to the geometry and energy minimization optimizes the altered structure. Homology modeling is a useful technique when a target protein of known sequence is related to at lease one other protein of known sequence and structure. If the proteins are closely related, the known protein structure (the parent) can serve as the basis for a model of the target.

The understanding is that functionally analogous proteins with homologous sequences will have closely related structures with common tertiary folding patterns. When sequence homology with a known protein is high, modeling of an unknown structure by comparison can be carried out with reasonable success. However, it is noted that structure homology may remain significant even if sequence homology is low, i.e. a 3D structure seems better conserved than the residue sequence. It is shown that protein pairs with a sequence homology greater than 50% have 90% or more of the residues within a structurally conserved common fold (Stewart *et al.*, 1987). The homology modeling normally consists of the following steps:

1. Start from the known sequences and align the sequences of the target and the protein or proteins of known structure. Often the InDels (insertions and deletions) between the related structures occur in the loop regions between α-helices and β-strands.

2. Assemble fragments/substructures from different, known homologous structures. Overlapping the main-chains of the target protein and the structure of the closely related protein of known structure.

3. Determine main-chain segments to represent the regions containing InDels. Stitching these regions into the main-chain of the known protein to construct a model for the complete main-chain of the target protein. Carry out limited structural changes from a known neighboring protein.

4. Retain the side-chain conformations for residues that have not mutated and replace the side-chains of residues that have been mutated.

5. Examine the model to detect any serious collisions between atoms.

6. Optimize/refine the model by energy minimization.

In principle, predicting structure from the sequence by comparison to a known homologous structure is satisfactory when sequence homology is greater than 50% (Chothia and Lesk, 1986). Part of the problem of homology modeling at lower levels of similarity is to correctly align unknown and target proteins. Sequence alignments are more or less straightforward for levels of above 30% pair-wise sequence identity. The region between 20% and 30% sequence identity (the twilight zone) is less certain. A means to automatically intrude into the twilight zone by detecting remote homologues (sequence identity <25%) are threading techniques (Bryant and Altschul, 1995). A sequence of unknown structure is threaded into a sequence of known structure and the fitness of the sequences for that structure is assessed. SWISS-MODEL at http://www.expasy.org/swissmod/SWISS-MODEL.html (Schwede *et al.*, 2003) and CaspR at http://igs-server.cnrs-mrs.fr/Caspr/ (Claude *et al.*, 2004) provide the facilities for automatic homology modeling.

16.5.7 *Ab initio* prediction of protein structure

It is well established that the sequence of the amino acid constituting a protein is of prime importance in the determination of its 3D structure and functional properties. Because sequence assignment is much faster and easier than 3D structure determination, the prediction of the 3D structure from the sequence of amino acids has been a great challenge for protein modeling. Energy-based structure prediction, known as the *ab initio* method (Bonneau and Baker, 2001) relies on energy minimization and molecular dynamics. The method is faced with the problem of a large number of possible multiple minima, making the traversal of the conformational space difficult, and making the detection of the real energy minimum or native conformation uncertain (subsection 9.2). PROTINFO at http://protinfo.compbio.washington.edu/ (Hung and Samudrala, 2003) offers the *de novo* method for modeling the protein structure consisting of less than 100 amino acid residues.

The genetic algorithm (Dandekar and Argos, 1994; Koza, 1993) is applied to the problem of protein structure prediction with a simple force field as the fitness function to generate a set of suboptimal/native-like conformations. Because the number of probable conformations is so large (for the main chain conformations with 2 torsion angles per residue and assuming 5 likely values per torsion angle, a protein of medium size with 100 residues will have $(2 \times 5)^{100} = 10^{100}$ conformations even if we further assume optimal (constant) bond lengths, bond angles and torsion angles for the side chains), it is computationally impossible to evaluate all the conformations to find the global optimum. Various constraints and approximations have to be introduced. For example, an assumption of constant bond lengths and bond angles by carrying out folding simulation in vacuum simplifies the total energy expression to

$$E = E_{tor} + E_{vdW} + E_{elec} + E_{pe}$$

where E_{tor}, E_{vdW}, E_{elec} and E_{pe} are torsion angle potential, van der Waals interaction, electrostatic potential and pseudo entropic term that derive the protein to a globular state ($E_{pe} \approx 4^{\Delta d}$ kcal/mol where Δd = largest distance between any $C\alpha$ atoms in one conformation). The combination of heuristic criteria with force field components may alleviate the inadequacy in the simplified fitness functions. The secondary structure prediction may be performed to reduce the search space. Thus either idealized torsion angles or boundaries for torsion angles according to the predicted secondary structures can be used to constrain main chain torsion angles.

It is shown that the incorrect structures have less stabilizing hydrogen bonding, electrostatic and van der Waals interactions. The incorrect structures also have a larger solvent accessible surface, and a greater fraction of hydrophobic side chain atoms exposed to the solvent.

16.5.8 Solvation

Solvation can have a profound effect on the results of molecular calculation. Solvent can strongly affect the energies of different conformations. It influences the hydrogen bonding pattern, solute surface area and hydrophilic/hydrophobic group exposures of protein molecules. Solvation is an important, ongoing problem in protein modeling. Two approaches are available to account for the solvent environment in protein modeling.

1. *Explicit solvent model*: The easiest way of incorporating discrete solvent is to include only a shell of solvent molecules surrounding the solute in molecular simulations. With this method, the solute-vacuum interface is replaced by a solvent-vacuum interface. A more realistic model involves placing the solute in a solvent box with periodic boundary conditions (Jorgensen *et al.*, 1983) where solvent molecules leaving one side of the box reappear at the opposite side so that the density of the system is maintained at a constant value. The dielectric constant defines the screening effect of solvent molecules on nonbonded (electrostatic) interactions and can vary from 1 (*in vacuo*) to 80 (in water). For the binding site, a constant dielectric of 4.0 is often used where no measured data are available. An explicit representation of the surrounding solvent can provide an accurate treatment of solute-solvent interactions, but it typically increases the system size. Furthermore, interactions with explicit solvent need to be average over relatively long time.

2. *Implicit solvent model*: Alternatively implicit solvent models yield significant computation efficiency by reducing explicit solute-solvent interactions to their mean field characteristics, which are expressed as a function of the solute configuration alone. These approaches generally reduce the time for a single energy calculation and avoid the need for averaging over the solvent degrees of freedom. An early simple model uses a distance-dependent dielectric constant to mimic the effect of solvent in MM calculations. Recent development of implicit solvent tools involves generalized Born approaches, dielectric screening function formulations and models based on solvent-accessible surface areas (Feig and Brooks, 2004).

16.5.9 On-line protein structure prediction

Most of online structure prediction servers apply homology modeling to construct 3D models. Table 16.12 lists useful on-line servers and tools for predicting/modeling protein structures.

The 3D-PSSM (Kelley *et al.*, 2000) server at http://www.bmm.icnet.uk/~3dpssm/ offers online protein fold recognition. The query sequence is used to search the Fold library for homologues. The submitter will be informed of the URL where the result is located for 4 days. The output includes a summary table (hits with statistics, models that can be viewed with RasMol, classifications and links) and fold recognition by 3D-PSSM. The FUGUE site (Shi *et al.*, 2001) at http://www-cryst.bioc.cam.ac.uk/~fugue/prfsearch.html is a protein fold recognition server. The query sequence is used to search structurally homologous proteins in the Fold library for hits (representatives). The submitter will be informed with a short summary of the result and the URL where the detailed result can be examined for 5 days. The fold recognition result are summarized in a view ranking listed according to Z-score (>=6.0, 5.0, 4.7 and 3.5 for certain, likely, marginal and guess with 99%, 95%, 90% and 50% confidence). The keys linking to the files for alignments

TABLE 16.12 Web servers for protein structure predictions

Web server	URL	Description	Ref
3D-PSSM	http://www.bmm.icnet.uk/~3dpssm/	Fold recognition	1
FUGUE	http://www-cryst.bioc.cam.ac.uk/~fugue/prfsearch.html	Fold recognition	2
CaspR	http://igs-server.cnrs-mrs.fr/Caspr/	Homology modeling	3
GeneSilico	http://genesilico.pl/meta	FR to 3D models	4
ModBase	http://alto.compbio.ucsf.edu/modbase-cgi/index.cgi	Annotated comparative 3D structures/models	5
PROSPECT-PSPP	http://csbi.bmb.uga.edu/protein_pipeline	FR to 3D models	6
PROTINFO	http://protinfo.compbio.washington.edu/	CM, *de novo* prediction	7
SA-Search	http://bioserv.rpbs.jussieu.fr/cgi-bin/SA-search	Protein structure mining	
SWISS-MODEL	http://swissmodel.expasy.org	Homology modeling	8

Note: References cited are: [1] Kelly *et al.* (2000); [2] Shi *et al.* (2001); [3] Claude, *et al.* (2004); [4] Kurowski, *et al.* (2003); [5] Sánchez *et al.* 3; [6] Guo, *et al.* (2004); [7] Hung and Samudral (2003; [8] Guex *et al.* (1999).

with structures/fold assignments are available from the View Alignments table. The query sequence is aligned against the sequence(s)/structure(s) of either all representatives or a selection of the group. Sequences of the representative proteins are displayed in JOY protein sequence/structure representation (http://www-cryst.bioc.cam.ac.uk/~joy/). GeneSilico (http://genesilico.pl/meta) is another fold recognition server. Fold recognition results from target-template alignments are converted into 3D models. PROSPECT-PSPP (http://csbl.bmb.uga.edu/protein_pipeline) provides multiple computational tools for the fully automated protein structure prediction. These include:

- sequence analysis including signal peptide prediction, protein type prediction and domain partition;
- secondary structure prediction;
- fold recognition; and
- atomic structural model generation.

ModBase (http://guitar.rockefeller.edu/modbase/) is a database of annotated comparative protein structure models. Since structure is more conserved in evolution than sequence, these models apply in the construction of all-atom 3D models in the comparative or homology modeling. CaspR applies the principle of molecular replacement (Šali and Blundell, 1993) to homology modeling and is accessible at http://igs-server.cnrs-mrs.fr/CaspR/. CaspR provides a progress report of the hierarchically organized summary sheets that describe the different stages of the computation with an increasing level of detail. The ten highest-scoring potential, pre-refined models are available for download in PDB format (Claude *et al.*, 2004).

SWISS-MODEL is the most popular online homology modeling server (Guex *et al.*, 1999; Schwede *et al.*, 2003) which can be access at http://swissmodel.expasy.org. The server offers three modes of user input:

1. *First approach mode*: This mode is for a simple interface that requires only an amino acid sequence as input data. The server automatically selects suitable templates and the modeling procedure starts if at least one modeling template with a sequence identity of more than 25% are available.
2. *Alignment mode*: In this mode, the modeling procedure is initiated by submitting a sequence alignment. The user specifies which sequence is the target sequence and which one corresponds to a structurally known protein from the ExPDB template library. The server constructs the model based on the given alignment.
3. *Project mode*: The project mode allows the user to submit a manually optimized modeling request. The starting point for this mode is a DeepView project file, which contains the superposed template structures and the alignment between the target and the templates. This mode gives the user control over a wide range of parameters and can be used to iteratively improve the output of the first approach mode.

The procedure involves template selection, target-template alignment, model building and energy minimization/evaluation. These steps are iteratively repeated until a satisfying model is achieved. A project of SWISS-MODEL, DeepView (Swiss-PDB Viewer) is an integrated application program that provides a user interface for visualization, analysis and manipulation of sequence-to-structure workbench. The program is available from the ExPASy at http://www.expasy.org/spdv. The PC version can be downloaded from http://www.expasy.org/spdbv/text/getpc.htm and used for standalone modeling (Tsai, 2002) for which an accompanying Spdbv Loop Database should be downloaded and placed into the '-stuff_' directory. The user guide for Spdbv is available from http://www.expasy.org/spdbv/mainpage.html

PROTINFO (http://protinfo.compbio.washington.edu) implements three methods, namely comparative modeling, fold recognition and *de novo* prediction to perform protein structure predictions (Hung and Samudrala, 2003). Other modeling servers dedicated to specific structural features of protein molecules are listed Table 16.13.

16.5.10 Protein–protein interaction

Recent developments in experimental and computational techniques allow a shift in the focus from the structures of individual proteins to their functions and therefore the protein–protein interactions (Waksman, 2005). Publicly accessible databases and servers (Table 16.14) of protein–protein interactions simplify the analysis, validation and understanding of protein interactions.

TABLE 16.13 Modeling servers for specific structure features

Server	URL	Structure feature
CASTp	http://cast.engr.uic.edu	Surface topology
CKAAPS DB	http://ckaaps.sdsc.edu/perl/browser.pl	Conserved key aa positions analysis
DSDBASE	http://ncbs.res.in/~faculty/mini/dsdbase/dsdbse.html	Disulfide bonds
GlobPlot	http://globplot.embl.de	Globularity and disorder
Molmov DB	http://www.molvdb.org	Macromolecular movements
POPS	http://mathbio.nimr.mrc.ac.uk/~ffranc/POPS	Solvent accessible surface area
pvSOAR	http://pvsoar.bioengr.uic.edu/	Amino acid surface patterns
Qgrid	http://www.netasa.org/ggrid/index.html	Charged/hydrophobic regions
SCit	http://bioserv.rpbs.jussieu.fr/SCit	Side chain conformation analysis
SURFACE	http://cbm.bio.uniroma2.it/surface	Surface res and function annotation

Note: Abbreviations used: aa, amino acids; res, residue(s).

TABLE 16.14 Online protein interaction databases and servers

Web site	URL	Description
ADVICE	http://advice.i2r.a-star.edu.sg	Postulation of PPI
BIND	http://bind.ca	Archives for cplext, pathway
ClusPro	http://nrc.bu.edu/cluster	Protein-protein docking
DIP	http://dip.doe-mbl.ucla.edu	Data collection/validation of PPI
PREDICTOME	http://predictome.bu.edu	Compilation of P&E data
STRING	http://www.bork.embl-heidelberg.de/STRING	Compilation of P&E data
InterWeaver	http://interweaver.i2r.a-star.edu/sg	Potential protein interactions
NCI	http://www.mrc-lmb.cam.ac.uk/genomics/nci/	Non-canonical interaction
iMOT	http://cps.imot-messen.in/imot/iMOTserver.html	Interacting motifs
iSPOT	http://cbm.bio.uniroma2.it/ispot/	Interaction specificity
PathBLAST	http://www.pathblast.org/	Align protein intnw
WebInterViewer	http://interviewer.jnha.ac.kr/	Analysis of molecular intnw
SCOPPI	http://www.scoppi.org	Classification of PPI interfaces
ProteMiner-SSM	http://p4.sbl.bc.sinica.edu.tw/proteminer/	Protein-ligand interaction
SiteBase	http://www.bioinformatics.leeds.ac.uk/sb	Protein-ligand interaction site

Note: Abbreviations used: cplext, complexation; PPI, protein-protein interaction(s); intnw, interaction network; P&E, prediction and experimental.

A structural description of the protein interactions or interaction networks of the genome complement (interactomes) is an important step toward understanding cellular processes and functional proteomes. Several experimental techniques (Piehler, 2005) can provide structural information about protein interactions. This information may be used to infer the configuration of interactomes. These techniques include two hybrid system (Stagljar and Field, 2002), tagged affinity chromatography (Ranish *et al.*, 2003), site-directed mutagenesis (Cunningham *et al.*, 1989), footprinting (Anand *et al.*, 2003), chemical cross-linking (Trester-Zedlitz *et al.*, 2003), novel spectroscopic method (Haustein and Schwille, 2004) and computational approach (Salwinski and Eisenberg, 2003). ClusPro at http://nrc.bu.edu/cluster/ (Comeau *et al.*, 2004) screens PDB database for the favorable surface complementarity to the input protein structures and selects those with good electrostatic and desolvation free energies for cluster analysis. The output of a short list of putative complexes ranked according to their clustering properties are returned.

With known 3D structures of proteins involved in the interaction, computational methods are available that suggest the structure of the interaction (Gray *et al.*, 2003). Most of docking methods aim to predict the atomic model of a complex by maximizing the shape and chemical complementarity between a given pair of interacting proteins (Gray *et al.*, 2003; Smith and Sternberg, 2002). Although docking methods are not sufficiently accurate to predict whether or not two proteins actually interact with each other, they can sometimes correctly identify the interacting surfaces between two structurally-defined subunits. Docking methods are also systematically assessed by blind trials in the Critical Assessment of Predicted Interactions (CAPRI) (Mendez *et al.*, 2003).

16.6 INVESTIGATION OF PROTEOME EXPRESSION AND FUNCTION

16.6.1 Two-dimensional gel electrophoresis

The readily available experimental tools for measurement of protein expression by two-dimensional electrophoresis (2DE), and for protein identification and characterization by mass spectrometry (MS) have made a significant impact on proteomics (Hamdan and Righetti, 2005). The coupling of 2DE and MS offers an efficient tool for investigating proteome expression.

The characterization of complex proteomes requires separation, detection and analysis of many thousands of proteins from whole cells, tissues or organisms. Cellular/tissue proteins are solubilized (Herbert, 1999) prior to their separation/analysis. Two-dimensional gel electrophoresis has demonstrated the potential to separate several thousand of proteins (Klose 1999; Rabilloud, 2000) in single experiment and has been the method of choice for the separation/purification in proteomic studies.

The application of the isoelectric focusing (IEF) technique coupled with sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) results in a 2D method that separates proteins according to two independent parameters, i.e. charge (isoelectric pH, *p*I) in the first dimension by IEF and size (relative molecular mass, M_r) in the second dimension by SDS-PAGE. The development of immobilized pH gradient (IPG) IEF based on the use of the Immobiline reagents greatly improve the desirability of 2DE in proteomic applications (Rabilloud, 2000; Görg *et al.*, 2004). 2DE has the capacity to support analysis of protein expression and post-translational modification (Figure 16.5). Owing to the high dynamic range and diversity of the proteins expressed in eukaryotic tissues, it is

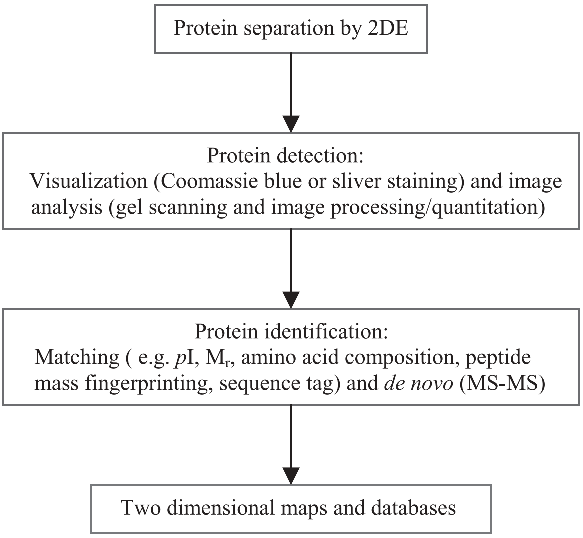

Protein separation by 2DE
Protein detection:
Visualization (Coomassie blue or sliver staining) and image analysis (gel scanning and image processing/quantitation)
Protein identification:
Matching (e.g. *p*I, M_r, amino acid composition, peptide mass fingerprinting, sequence tag) and *de novo* (MS-MS)
Two dimensional maps and databases

Figure 16.5 Scheme for proteome expression analysis by two-dimensional electrophoresis Cell/tissue proteins are separated by 2DE, the combination of IEF in the first dimension and SDS-PAGE in the second dimension. Coomassie blue and silver stains are the commonly used visualization methods for routine detection and quantitation of separated proteins. The scanning process converts the analog gel image into a digital representation for computer-based processing. Each detected protein spot is assigned molecular mass (M_r) and isoelectric point (*p*I). Thus the numbering (location) of the protein spot is based on the M_r/*p*I grid using some characteristic protein spots (reference spots) for which M_r/*p*I values are known. The M_r/*p*I values can be compared to/obtained from 2DE databases, e.g. 2DE server at ExPASy (http://www.expasy.ch/ch2d/). The identification of separated proteins form 2DE is coupled to mass fingerprint using MALDI-TOF MS or sequence analysis by MS-MS.

sometimes necessary to perform a prior fractionation step to reduce the complexity of the sample and enrich for low copy number proteins.

The application of 2DE in proteomics depends on its resolving capacity and reproducibility. The resolution of 2DE is dependent on the separation length in both dimensions and is usually taken to be proportional to the total gel area available for the separation. For example, a standard 18-cm long IPG IEF gels in combination with 20-cm long SDS-PAGE gels, can separate ~2000 proteins from complex mixtures of whole cell/tissue lysates. For maximal resolution of complex protein mixtures, larger formal gels may be used (Klose, 1999). The reproducibility is essential for the wide use of 2DE databases to compare 2DE separation patterns generated in different laboratories as well as interlaboratory studies of various types of cellular/tissue samples or samples of different physiological states. In two-dimensional difference in-gel electrophoresis (2D-DIGE), which is used to analyze differentially expressed proteins from two different cell populations (Tonge *et al.*, 2001), one sample is labeled with one of the dyes (e.g. red fluorescent Cy3) and other sample with the other dye (e.g. green fluorescent Cy5). The two samples are mixed and run on the same 2DE. Identical proteins from each pool migrate to exactly the same position. For example, by using red (Cy3) and green dye (Cy5) dyes, the proteins that are expressed equally (i.e. equal concentrations) in the two samples become yellow. Those upregulated in one of the samples are red or reddish and those upregulated in the other are green or greenish. It should be remembered that labeled proteins are 0.5 kDa higher in molecular mass than the unlabeled resulting in different positions of labeled and unlabeled proteins. 2DE has full sequence coverage, giving it the ability to monitor unknown post-translational modifications that change the migration of proteins.

The protein spots can be localized and identified in federated 2DE databases (Appel *et al.*, 1994). The visual interlaboratory comparison of two gel images featured in a 2DE metadatabase is realized by a Flicker-server (http://www-lmmb.ncifcrf.gov/flicker) in which two images (source and target images) can be visualized alternatively. The Web accessible 2DE databases are indexed in the WORLD-2DPAGE index (http://www.expasy.ch/ch2d/2d-index.html). The use of dedicated equipment such as DALT (Amersham Pharmacia) and Investigator (Genomic Solution), which permit the simultaneous electrophoresis of large numbers of 2DE gels under reproducibly controlled conditions, result in 2DE protein separations with high spatial and quantitative reproducibility (Bloomberg *et al.*, 1995). For proteome expression profiling, 2DE is coupled to mass spectrometric identification of the separated proteins.

16.6.2 Proteome analysis by mass spectrometry

Since the introduction of electrospray ionization (ESI) (Fenn *et al.*, 1989) and matrix-assisted laser desorption ionization (MALDI) (Beavis and Chait, 1996; Kaufmann, 1995), which permit the efficient transfer of biomacromolecules into the gas phase, mass spectrometry (MS) has become the method of choice for carrying out internal sequencing of peptides/proteins in proteomics (Aebersold and Goodlett, 2001; Mann *et al.*, 2001; Reinders *et al.*, 2004). The four primary advantages of MS sequencing include the high sensitivity, the rapid speed of the analysis, the large amount of information generated in each experiment, and the ability to characterize post-translational modifications. Routine applications of MS can detect and sequence peptides of femtomole levels. The key consequence of this sensitivity is that sequencing and identification can be accomplished on the same sample used in routine experiments (Figure 16.6).

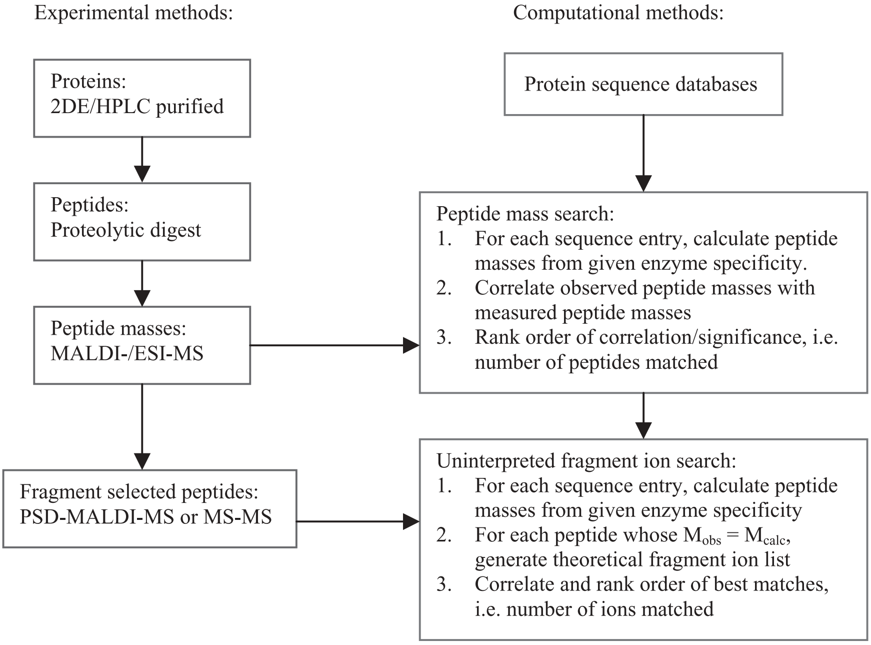

Experimental methods:
Computational methods:
Proteins:
2DE/HPLC purified
Protein sequence databases
Peptides:
Proteolytic digest
Peptide mass search:
1. For each sequence entry, calculate peptide masses from given enzyme specificity.
2. Correlate observed peptide masses with measured peptide masses
3. Rank order of correlation/significance, i.e. number of peptides matched
Peptide masses:
MALDI-/ESI-MS
Uninterpreted fragment ion search:
1. For each sequence entry, calculate peptide masses from given enzyme specificity
2. For each peptide whose $M_{obs} = M_{calc}$, generate theoretical fragment ion list
3. Correlate and rank order of best matches, i.e. number of ions matched
Fragment selected peptides:
PSD-MALDI-MS or MS-MS

Figure 16.6 Protein identification by mass spectrometry

Protein identification by MS relies on the generation of information unique to a protein and the availability of correlating information in the form of protein or DNA sequences in proteomic/genomic databases. It was realized that the accurate measurement of a protein mass was insufficient for identification by protein database searches. However, the accurate masses of peptides, generated by proteolytic digestion of a protein were often sufficient to identify unambiguously the protein in databases. MALDI-TOF MS becomes the method of choice to rapidly generate mass fingerprints of proteolytic digests. Sequence information for MALDI is obtained with approaches for either studying post-source decay reactions or enzymatic protocols to create a nested set of fragments. The high sensitivity is achieved through the use of a time-of-flight (TOF) mass analyzer for the second mass analyzer, which has the added effect of giving relatively high m/z resolution. MALDI is able to generate high mass ions and its pulsed ion generation is ideal for TOF mass analysis with the development of MALDI-TOF. Increased mass accuracy m/z analysis of MALDI-TOF is primarily a result of better resolution by the use of one of the two approaches; the use of ion reflectors or reflectrons (Cornish and Cotter, 1997), and the use of the technique of time-lag focusing or delayed ion extraction (Brown and Lennon, 1995). Current TOF mass analyzers generally use a combination of both delayed ion extraction and reflectron ion optics to improve the accuracy of peptide mass measurement, which has dramatically improved the effectiveness of peptide mass-mapping experiments (Jensen *et al.*, 1997). Amino acid sequence information can be obtained from peptide ions formed by MALDI if the peptide ions can be fragmented. These fragment ions are produced by fragmentation reactions that occur at various times after ionization and therefore are often referred to as post-source decay (PSD) reactions. One method for proper mass measurement of these reaction products is to delay ion extraction into the TOF mass analyzer long enough to allow the reaction to take place. This approach has often been applied to the identification of 2DE separated proteins in proteome expression profiling. The software to search protein/DNA databases using peptide mass fingerprints is available at http://www.expasy.ch/tools/

As mentioned earlier (subsection 4.2), it is common to produce multiply charged analytes by ESI when dealing with peptides and proteins. However, ESI-MS has two advantages. First, they can be easily coupled to different sample separation and sample introduction techniques. The second advantage is the increase in the quality of the tandem mass spectra generated from multiply charged analytes. ESI is a soft-ionization technique so that little fragmentation, and therefore little structural information is seen in electrospray mass spectra. As a result, ESI depends on the combination with tandem MS to give sequence information. In tandem mass spectrometry (MS–MS), mass analysis is performed in two stages. In the first stage, a specific ion of interest is selected according to its mass from the set of ions that constitute the conventional mass spectrum. The second stage generates, from its selected precursor ion, a spectrum of product ions that arise from metastable ion (MI) or collision-induced dissociation (CID). Thus MS–MS provides a means for fragmenting a mass-selected ion and measuring the m/z of the product ions that are produced in the collision-induced fragmentation of ions (Kinter and Sherman, 2000).

MS–MS can provide enhanced information on the individual peptide contained in a proteolytic digest, facilitating the identification of the proteins and offering the possibility of *de novo* sequencing (Kinter and Sherman, 2000) when no representative entry is in a database. Typically in a first pass, the introduced peptides are separated according to m/z by the mass spectrometers (MS spectrum). A list of peptides with signals above a pre-established threshold is created. In a second pass, a mass window centered on a selected peptide is isolated by the mass spectrometer and the kinetic energy of the selected peptide is increased. The collision of the peptide with small gas molecules (CID) transfers sufficient energy to break the peptide bond(s) generating charged and neutral fragments. Then in the third pass, the generated charged fragments are separated by the mass spectrometer according to their m/z, creating the MS–MS spectrum. Because the fragmentation occurs at the peptide bonds, a ladder of fragments is generated and the mass difference between two fragments of the same type in the ladder corresponds to an amino acid. The sequence of the peptide can then be reconstituted by a fragment-walk along the ladder. The generation of MS–MS spectra is performed in a serial manner allowing only one peptide MS–MS spectrum to be acquired at a time. Hence the coupling of separation techniques or low flow infusion techniques and MS–MS via the ESI interface, allows efficient serial analysis of peptides form peptide mixtures to be performed. In general, peptide mass-mapping experiments are the most common application of ESI-CID to proteomic research. These experiments do not require any amino acid sequence data *per se* but rather identify the source protein with specific, informative molecular weight information to reveal the amino acid sequences of the peptide in the digest. Currently, the *de novo* sequencing from MS–MS spectra is essentially a manual technique. Algorithms that search uninterpreted MS–MS spectra against protein/DNA databases are available and can be accessed at a number of Web sites (Table 16.15), which provide peptide mass and fragment ion search programs.

Additional analyses known as orthogonal methods have been developed to increase the confidence of peptide-mass assignments. The orthogonal information restricts the number of peptides that match the isobaric (same mass) peptides, which are identical in the original database search based on peptide masses. The orthogonal methods include the following categories:

- *Site-specific chemical modifications*: For examples, methyl esterification that adds 14 m.u. for each carboxyl group (i.e. side chains of Asp, Glu or C-terminus in the peptide); iodination that adds 126 m.u. for each Tyr.
- *Determination of partial amino acid composition of peptides*: This can be accomplished by the identification of immonium ions ($H_2N{=}CHR^+$) from the MS–MS spectra (Table 5.3). Alternatively, an exchange experiment with deuterium may increase 0 to 5 m.u. per residue for the exchangeable hydrogens in a peptide. The total increase in peptide mass therefore reflects the peptide composition.
- *Identification of the N-terminal and/or C-terminal amino acid residues*: The N-terminal amino acids of peptides are determined chemically by one step of Edman degradation or enzymatic removal by the use of an aminopeptidase. Similarly the

TABLE 16.15 Resource sites with MS-based protein identification tools

Resource	URL
CBRG	http://cgrg.inf.ethz.ch/MassSearch.html
Proteome, ExPASy	http://www.expasy.ch/tools/#proteome
Mascot	http://www.matrix-science.com/
Peptide Search, EMBL	http://www.mann.embl-heidelberg.de/Services/PeptideSearch/
ProSight PTM	http://prosightptm.scs.uiuc.edu/
ProteinProspector, UCSF	http://prospector.ucsf.edu/
Rockefeller Univ.	http://prowl.rockefeller.edu/
SEQNET	http://www.seqnet.dl.ac.uk/Bioinformatics/welapp/mowse
SEQUEST	http://thompson.mbt.washington.edu/sequest/

C-terminal amino acids can be determined by the use of carboxypeptidase. The subtractive mass and the knowledge of the N- and/or C-terminal residues provide the information of the peptide identification.

- *Identification of different cleavage sites within the peptides*: The determination of the presence and relative location of a specific residue in a peptide, either secondary digestion of the primary digest with an enzyme of differing specificity or parallel digestions of aliquots of the same sample with enzymes of different specificity can be employed.

In general, the intensity of a peptide ion signal is taken to reflect the quantity of that peptide present. However, isotope-coded affinity tag (ICAT) technique (Gygi *et al.*, 1999) may improve the quantitative characteristics of MS. The technique implies that peptides of identical chemical nature, only differing in mass because of differences in isotopic composition, are expected to produce identical signals in MS. For example, samples are labeled with an alkylating group (e.g. iodoacetate), which covalently attached to Cys in the protein. This is coupled to a polyether linker and a biotin affinity tag. The linker may contain eight hydrogen atoms (light sample) or eight deuterium atoms (heavy sample). The samples are combined and subjected to enzymatic digestion. The ICAT-peptides are then enriched by avidin (specific for biotin) affinity chromatography and subsequently analyzed by MS–MS. Ideally each cysteinyl peptide will appear as a pair of signals that differ by the mass difference of the mass tag (i.e. 8 Da), if only one Cys is present in the peptide. Thus the ratio between the two signals will reflect the ratio between the proteins in the samples from where peptides are obtained.

16.6.3 Analysis of posttranslational modification by mass spectrometry

Significant sequence information about proteomes can be derived from genomic studies. However, various posttranslational modifications (subsection 13.5.5), which affect structures, properties and functions of proteins cannot be derived or predicted accurately from genomics. Proteomic techniques are the only solution to investigate posttranslational modifications (PTMs) unambiguously. Protein Modification Screening Tool (ProMoST) at http://proteomics.mcw.edu/promost offers a web-based tool for mapping protein modifications on 2D gels (Halligan *et al.*, 2004). Mass spectrometry in combination with affinity-based enrichment/separation offers an attractive avenue for systematic investigation of posttranslationally modified proteins in proteomics (Jensen, 2004). Among those many types of PTMs, few have been shown to be commonly occurring modifications that affect a broad range of biological function and processes. Protein glycosylation has been recognized as one of the most prominent biochemical alternations associated with pathological manifestations and will be dealt in the next chapter. Reversible protein phosphorylation is the most common PTM of great regulatory importance to many biological processes. It has been investigated extensively to understand enzymology of the protein kinases/phosphatases catalyzing phosphorylation/dephosphorylation, which affect the structure and function of the target proteins (Krebs and Beavo, 1979). The functional consequences of reversible protein phosphorylation provide insights into how biological systems are controlled at the molecular level. Mass spectroscopic studies of protein phosphorylation will be considered to illustrate an approach to analysis of PTM by MS.

The aims of studying protein phosphorylation are three-fold:

1. to determine the amino acid residues (sites) that are phosphorylated *in vivo* for a given protein in a specific state in a cell;

2. to identify the kinase for the phosphorylation and phosphatase for the dephosphorylation event;
3. to analyze the functional consequences of the observed phosphorylation events.

The first aim can be directly addressed by MS (Mann *et al.*, 2001) Initially phosphopeptides are separated/enriched by two-dimensional phosphopeptide (2D–PP), reverse-phase high performance liquid chromatography (rpHPLC) or immobilized metal affinity chromatography (IMAC) (Gatti and Traugh, 1999; Ficarro *et al.*, 2002) or immunoprecipitation using antiphosphotyrosine (anti-pTyr) or anti-phosphoserine/phosphothreonine (pSer/pThr) antibodies (Gronborg *et al.*, 2002). Mass spectroscopic analysis of phosphopeptide follows:

- *Phosphatase treatment*: The phosphopeptides in the mixture can be identified by recording MS (usually MALDI-MS) spectra before and after the treatment with alkaline phosphatase or phosphoprotein phosphatase. A net mass differential of 80 Da is caused by the addition of a phosphate group to Ser, Thr or Tyr residue. A prior knowledge of the amino acid sequence of the peptide facilitates the identification of the phosphorylation site.
- *Precursor ion scanning in the negative mode*: MS–MS in this mode (neutral or alkaline condition) scans the reporter ion for the phospho group, m/z = 79. Phosphopeptides are identified by observation of 97 ($H_2PO_4^-$), 79 (PO_3^-) and 63 (PO_2^-) Da. Once the phosphopeptide is identified, the reminder of the sample (after reconstitution in acidic solution) is analyzed in the conventional positive mode.
- *Phosphopeptides can be sequenced using MS–MS in the positive mode*: ESI in the positive mode (acidic condition) followed by in-source CID, the fragmentation spectrum will reveal a loss of 98 Da (H_3PO_4) due to β-elimination in the case of pSer and pThr, whereas pTyr is more stable. The location of pSer can then be identified either by a mass difference of two successive fragments of 167 (87 + 80) Da or 69 Da (β-elimination product, dehydroalanine). Similarly pThr can be located as a mass difference of 181 (101 + 80) Da or as 83 Da (β-elimination product, dehydroamino-2-butyric acid). The location of pTyr can be identified by a mass difference of 243 (163 + 80) Da.
- *Parent ion scanning in the positive mode for phosphotyrosines*: Taking advantage of the greater resolving power of mass spectrometer that can distinguish the characteristic reporter ion, the pTyr immonium ion at m/z = 216.04 from the amino acid doublets and triplet series (e.g. m/z = 216.09 of b_2 series of NT and QS doublets).

Other PTMs commonly found in proteins during routine proteomic analysis are ubiquitination, acetylation and N- or C-terminal processing. Protein ubiquitination plays an important role in regulation of protein degradation (subsection 12.8.2). Tryptic digestion of ubiquitinated proteins produces a signature peptide at the ubiquitination site containing a diGly remnant (G-G) which can be captured by the use of His-tagged ubiquitin and be detected by MS (Peng *et al.*, 2003). An analysis of acetylated peptides is useful because it helps in determining the amino terminus of mature proteins formed after PTM. Generally, all modifications that lead to a mass change can be analyzed by MS before and after de-modifications.

16.6.4 High throughput protein crystallography

The demand for a high-throughput protein crystallography (HTPC) is driven by the need to understand exactly how proteins bind to their ligands (Blundell *et al.*, 2002) in meeting the challenge of the genome-scale protein structure and function information (Pechkova and Nocolini, 2003). The bindings of high-affinity ligands (including biomacromolecules) to proteins often induce conformational changes. The knowledge on protein structural changes accompanying ligand/biomacromolecule interactions is essential in deciphering the molecular mechanisms of protein functions. The solution of a protein crystal structure is a time-consuming and complex process. HTPC generally involves five phases:

1. Expression of enough protein for analysis. In some cases, this step requires cloning or subcloning, transcription/translation.
2. Purification of the protein to yield stable protein preparation in a high concentration (e.g. in mg/mL range) suitable for crystallization.
3. Production of a sufficiently high-quality crystal by whatever means possible. Although many years of experience have led to a vast knowledge base and literature on finding the right condition to turn dynamic protein chains into crystal forms (Bergfors, 1999), this is arguably the toughest step. Biological Macromolecular Crystallization Database (BMCD) at http://www.bmcd.nist.gov:8080/bmcd/bmcd.html lists crystallization experiments that yield crystals from which a structure could be determined.
4. Collection of X-ray diffraction data. This is usually a skilled task performed by an X-ray crystallographer. The use of larger and better X-ray diffractomers, including those with synchroton sources, means that better data become available from smaller crystals. For example, a conventional single-crystal X-ray crystallography requires a crystal between 100–300 μm long, but bright focused X-rays can garner structural information from crystals of ~50 μm scale.
5. Conversion of X-ray data into a refined model of the protein structure using dedicated software. The aim of HTPC is to develop integrated software that will minimize human intervention in protein structure determination.

16.6.5 Protein–protein interactions by two-hybrid assay

In biological systems, proteins often function by interacting with other proteins. Identification and characterization of protein interaction and entire interaction networks (interactomes) is a prerequisite to understanding functional proteomes. Protein–protein interactions can be analyzed by various methods (Phizichy and Field, 1995; Piehler, 2005) among which the use of yeast two-hybrid assays (Lecrenier *et al.*, 1998) exploits the finding that eukaryotic transcription of genes to mRNA is controlled by transcription factors. Many transcription factors are composed of two domains that are physically separable and remain active provided they are in close proximity. The DNA-binding domain (BD) binds to the promoter region of the gene specific for the transcription factor, while the transcription activation domain (AD) in concert with the RNA polymerase recruits the rest of the element needed for the transcription (Figure 16.7A). In a typical two-hybrid experiment, a given protein (bait) is expressed as fusion with BD (bait fusion) (Figure 16.7B), whereas the putative interacting protein (prey) is fused to AD (prey fusion) (Figure 16.7C). The bait and prey carried by distinct plasmids, are co-expressed into yeast. If the two proteins interact, an active element consisting of both BD and AD are formed, and gene transcription can proceed (Fiure 16.7D).

The active element serves as a transcription factor for a reporter gene (e.g. β-galactosidase of lac*Z*) or prototropic marker (e.g. HIS*3*, ADE*Z*) that are under the control of upstream activating sequence (UAS) recognized by BD. Thus only those transformants

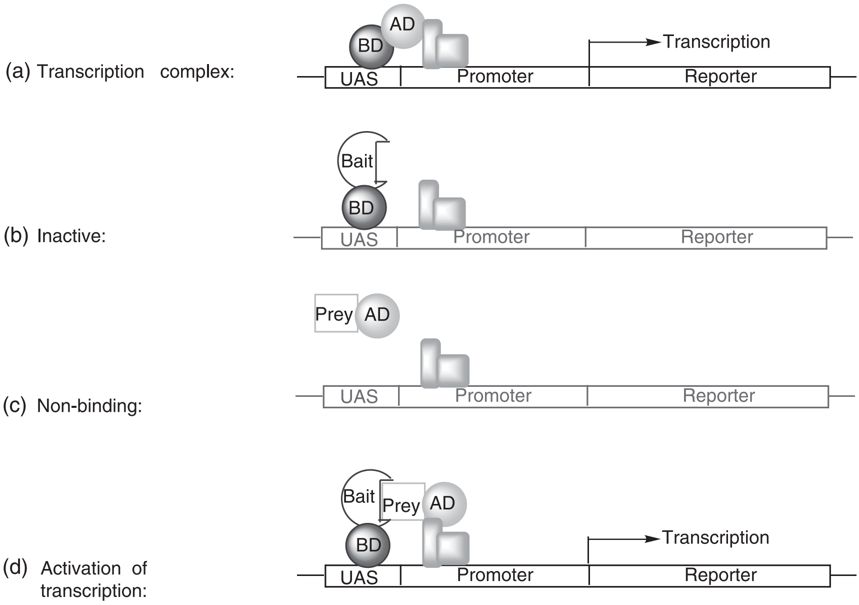


Figure 16.7 Schematic presentation of two-hybrid assay system
The transcription occurs when the transcription factor consisting of DNA-binding domain (BD) which binds at the upstream activating sequence (UAS) and transcription activation domain (AD) which recruits transcription elements and interacts with the promoter region (A). Bait fusion BD binds to UAS but cannot activate the transcription without AD (B). Likewise prey fusion AD cannot bind to UAS and does not activate transcription (C). The interaction of a bait and prey reconfigures the transcription element consisting of both BD and AD to activate transcription of the reporter gene. Adapted from Lecrenier, Foury and Goffeau (1998)

in which bait-prey interactions occur, will develop a blue color (lacZ reporter) in the presence of nitrophenyl-β-galactoside or grow on media lacking the selective His (HIS3) or adenine (ADEZ) markers.

It appears that the two-hybrid assays tend to identify a large number of false positives (Hengen, 1997). Furthermore, the conventional two-hybrid method detects interactions occurring in the nucleus using a transcriptional readout. Consequently the interaction of many membrane proteins and transcription factors cannot be measured. An approach especially useful for detecting interactions among membrane proteins is the split ubiquitin system (Stagljar *et al.*, 1998; Wittke *et al.*, 2002). The method involves bringing together two halves of ubiquitin in which the N-terminal fragment is fused to one protein (bait) and the C-terminal fragment is fused to a transcription factor and a putative interaction protein (prey). The interaction of the bait and prey proteins results in the interaction of the ubiquitin fragments and release of the transcription factor, which activates a reporter construct.

The two-hybrid system has been used to identify interactions among the entire complement of proteins encoded by the genomes (Uetz *et al.*, 2000). The two-hybrid system can be scaled and automated for the systematic search of protein–protein interactions. The data have proven to be valuable when integrated with protein–protein interaction data from various sources in interaction proteomics. The method can be adapted/modified to study protein–nucleic acid interactions (Putz *et al.*, 1996; Alexander *et al.*, 2001) and protein-ligand interactions (Licitra and Leu, 1996).

16.6.6 Protein chip

Currently, protein expression mapping is often performed by coupled 2DE or HPLC and MS–MS technologies. However, the development of protein microarrays (protein chips) may provide another powerful method to explore protein expression (protein profiling) and function on a genome-wide scale (Fung, 2004; LaBaer and Ramachandran, 2005; Hultschig *et al.*, 2006). Figure 16.8 shows some potential applications for protein microarrays.

The use of protein chips in functional proteomics has the following advantages:

- It can be scaled and automated for the high throughput process.
- Experimental conditions can be well controlled, such as the stringency of the binding/biological activities can be adjusted by the addition of cofactors or inhibitors.
- The microarray analyses are highly parallel and are not biased toward abundant proteins.
- It can be applied to identify the downstream targets of various enzymes/proteins by using proper detection methods.
- It can be used to identify *in vivo* posttranslational modifications such as glycosylation (with lectins) or phosphoryalation (with antibodies).
- It has a potential of permitting analysis of the kinetics of protein-protein/nucleic acids/glycan/ligand interactions *via* real-time detection methods.

Two general strategies have been pursued for the application of protein chips in proteomics:

1. Abundance-based (analytical) microarrays, which seek to measure the abundance of specific biomolecules (e.g. proteins) using analyte-specific reagents such as monoclonal antibodies. For this purpose, capture microarrays are generated by spotting specific capture

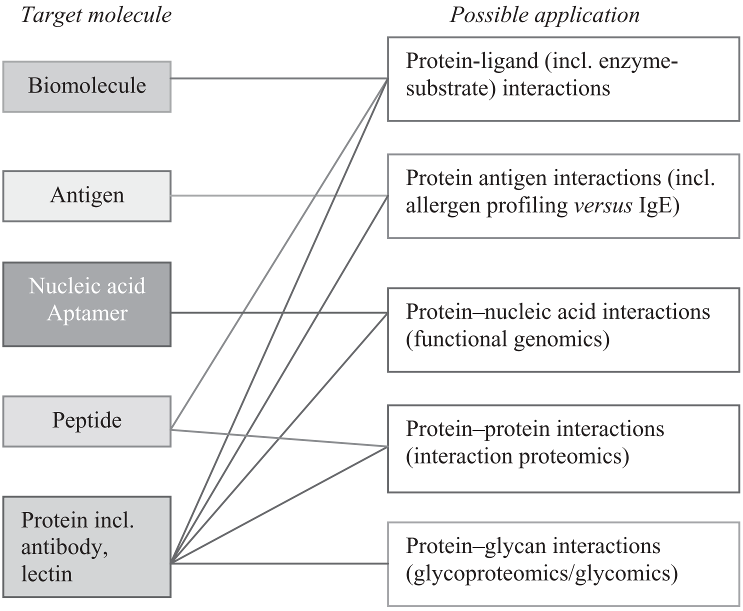

Target molecule
Possible application
Biomolecule
Antigen
Nucleic acid Aptamer
Peptide
Protein incl. antibody, lectin
Protein-ligand (incl. enzyme-substrate) interactions
Protein antigen interactions (incl. allergen profiling *versus* IgE)
Protein–nucleic acid interactions (functional genomics)
Protein–protein interactions (interaction proteomics)
Protein–glycan interactions (glycoproteomics/glycomics)

Figure 16.8 Potential applications of protein microarrays

molecules (i.e. analyte-specific reagents such as antibodies, aptamers) on the array surface to trap and assay their targets from complex mixtures (e.g. expressed protein mixtures). Typically, the capture microarray is probed with a complex sample, and then relative amounts of the targeted analytes can be determined by comparison to a reference sample. Capture microarrays have been used for comparative protein profilings of normal and diseased cells (Sreekumar *et al.*, 2001; Miller *et al.*, 2003).

2. Function-based (functional) microarrays which seek to study the biochemical properties and activities of target proteins printed on the arrays (Predki, 2004). These microarrays are produced by printing the proteins of interest on the array using methods designed to maintain the integrity and activity of proteins and allowing target proteins to be simultaneously screened for function (Jona and Synder, 2003; Mitchell, 2002). The function-based microarray can be used to examine protein interactions with other biomacromolecules and small biomolecules, construct protein interaction networks, probe enzyme activity and substrate specificity, screen for lead and candidate compounds in drug discovery, and integrate into metabolomics technology (Griffin and Shockcor, 2004; Saghatelian and Cravatt, 2005).

In functional protein microarrays, set of proteins of interest or an entire proteome is over expressed, purified and spotted in an addressable microarray surface. The most obvious target molecules that can effectively be used to investigate protein binding and therefore protein microarray analysis are monoclonal antibodies or IgG. However, unlike DNA microarrays, antibody arrays may pose some practical concerns:

- maintenance of the structural integrity and stability of antibodies during microarray fabrication;
- difficulty for finding experimental conditions that can be optimized simultaneously for all of the proteins being assayed because of the significant heterogeneity of antibody affinities;
- preservation of antigen epitopes and affinities during fluorescence labeling of proteins.

Approaches aimed at achieving quantitative protein profiling using antibody arrays in a multiplex format include signal amplification, multicolor detection and competitive displacement assays (Barry and Soloviev, 2004). If the use of antibodies for protein microarrays is to be materialized, several important criteria have to be fulfilled:

- access to a process that can support the generation and selection of antibodies to individual proteins and possibly each of their post-translationally modified forms;
- The antibodies must be highly specific, i.e. they must be capable of recognizing and distinguishing the individual protein when present in a complex mixture;
- The binding of the individual proteins within the mixture to their cognate antibodies must occur under similar conditions.

The ability to produce/isolate specific antibodies to a number of different proteins from the phage display library offer a potential to provide antibodies for genome scale application of protein chips (Lee and Mrksich, 2002). Phage display of combinatorial antibody library (subsection 13.7.5) is uniquely capable of providing specific monoclonal antibodies in the number and at the rate required for bioinformatics research. Table 16.16 lists some commercial sources of phage display antibody libraries.

Two schemes to affix proteins to the microarray surface are generally employed, chemical linkage and peptide fusion tags. In the chemical linkage format, proteins are

TABLE 16.16 Phage display antibody library

Source	Antibody	Web address
BioInvent	CoDeR-Fab	http://www.bioinvent.com
Cambridge Antibody Technol.	ScFv	http://www.cambridgeantibody.com
Crucell	HuNADE scFv	http://www.crucell.com
MorphoSys	HuCal*GoldFab	http://www.morphosys.com

immobilized to functionalized (amine, aldehyde, activated ester or epoxy) surfaces in random orientations using their carboxy, amino or thiol residues. Because the proteins are spotted close to the array surface, the protein folding and accessibility may be affected by this format. Peptide fusion tags (e.g. polyhistidine tag) can be appended to the amino or carboxyl terminus of the coding sequences for the target proteins. The resulting chimeric proteins are then immobilized via peptide tags. This format has the advantage that proteins are uniformly orientated at a distance from the array surface. For example, proteins are tagged with oligohistidine and the microarray surface is spotted with a nickel form of nitrilotriacetatic acid (six coordination sites of which four are occupied by Ni and two positions are available to bind His). The oligohistine-tagged proteins then form nickel complexes and are immobilized on the surface. A similar approach using a biotin tag on the streptavidin-derivatized surface binds biotin-tagged proteins to the streptavidin surface.

To circumvent the difficulty in obtaining purified proteins for functional protein microarrays, a different strategy, self-assembling protein microarray called nucleic acid programmable protein array (NAPPA) has been developed (Ramachandran *et al.*, 2004). In this method, full length cDNA molecules are immobilized on a microarry surface and expressed *in situ* using a mammalian cell-free expression system. A fusion tag present on the protein is recognized by a capture molecule arrayed along with the cDNA on the surface. This capture reaction immobilizes and arrays the proteins on the chip surface. Since both the target and query (probe) proteins are co-expressed, NAPPA eliminates the need to purify proteins and causes less concern about protein stability (LaBaer and Ramachandran, 2005).

16.6.7 Activity-based probe

The functional genomics based solely on genomic analyses, which place their reliance on mRNA levels (transcript profiling) as an indirect measure of protein quantity and function, is potentially risky because mRNA levels are generally poor predictors of protein abundance (Gygi *et al.*, 1999). Furthermore, proteins undergo post-translational modifications, which in turn affects their activities. Global measurements of protein abundance may again provide only an indirect assessment of protein function. Therefore functional genomics must approach from proteomics initiatives that are able to monitor, quantify and correlate protein abundance and activity. Chemical approaches offer important methods and reagents for the global analysis of protein function. For the activity-based proteomics, trifunctional molecules are synthesized as activity/affinity probes known as activity-based probe (ABP) or affinity-based probe (AFBP). The ABP requires active enzymes to facilitate covalent modification, whilst AFBP is designed to target noncatalytic residues on proteins and enzymes. Therefore AFBP requires highly selective tight binding to targets to be useful probes for distinct protein/enzyme families (Cravatt and Sorensen, 2000). In their most basic form, the activity probes consist of three distinct functional elements:

1. a reactive group for covalent attachment to the enzyme;
2. a linker that can modulate reactivity and specificity of the reactive group; and
3. a tag for identification (e.g. fluorescent or isotope label) or purification (e.g. affinity) of modified enzymes (Figure 16.9).

Such ABP generally meet the following criteria:

- React with a broad range of enzymes from a particular class directly in complex proteomes.
- React with these enzymes in a manner that correlated with their catalytic activities.
- Display minimal cross-reactivity with other protein classes.
- Possess a tag for the rapid detection and isolation of reactive enzymes.

Thus a successfully designed ABP permits the comparative measurement and identification/isolation of all of the active members of a given enzyme family present in two or more proteomes.

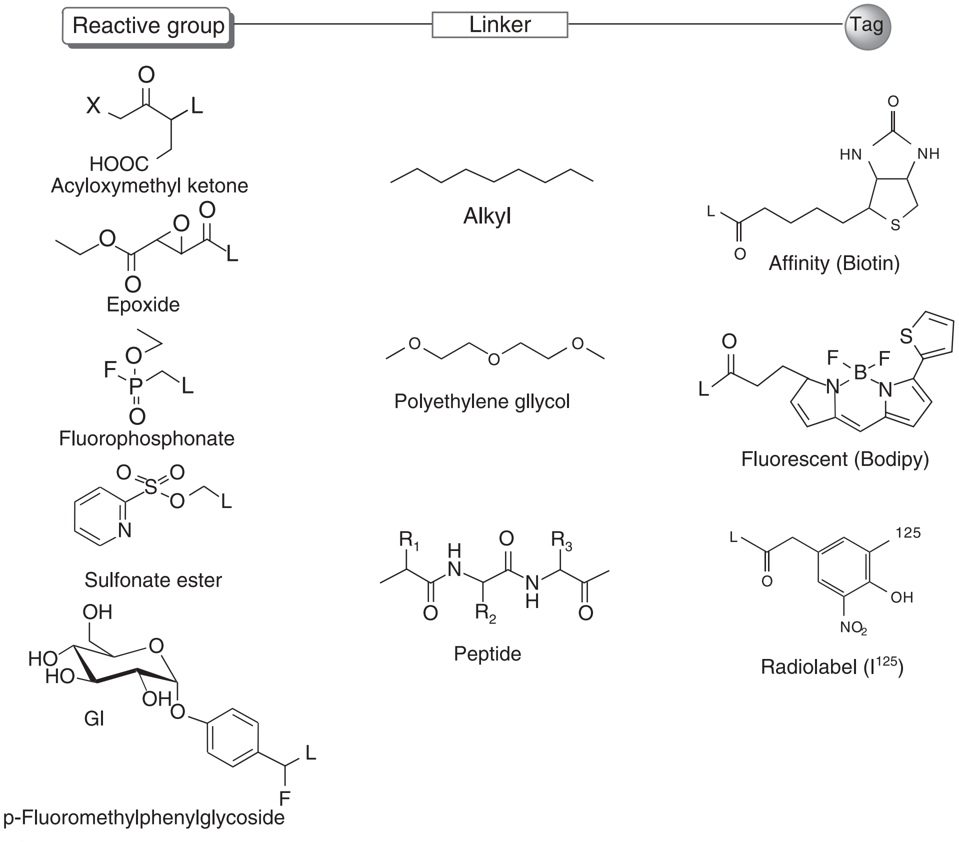

Reactive group
Linker
Tag
Acyloxymethyl ketone
Epoxide
Fluorophosphonate
Sulfonate ester
p-Fluoromethylphenylglycoside
Alkyl
Polyethylene gllycol
Peptide
Affinity (Biotin)
Fluorescent (Bodipy)
Radiolabel (I125)

Figure 16.9 Structures of chemical components of activity/affinity probes
The activity/affinity-based probe consists of reactive group and tag connected by a linker (L indicates the points of connection to the linker). Specific examples of each of the chemical components are shown. Two types of tags are commonly employed. The affinity tag for isolation/purification and the fluorescent or radiolabel tags for identification. Taken from Jeffery and Bogyo (2003)

16.6.7.1 Reactive group. The greatest challenge in the design of ABP/AFBP is selection of a reactive group because it must be both reactive toward a specific residue on a protein/enzyme and inert toward other reactive species within the cell/cell extract. The experiences gained from enzyme active-site studies by chemical modifications (subsection 11.4.2 and Table 11.9) provide an entry point to the selection of reactive groups. Knowledge on mechanistic differences of individual enzyme families further improves the choices. There are four general types of reactive groups that have been used to design chemical probes:

1. *Mechanism-based probes*: Knowledge of enzyme reaction mechanisms permits selection based on the differential reactivities of the catalytic nucleophiles used by individual enzyme subfamilies. For example, ABP bearing an acyloxymethyl ketone halide has been used to profile the caspases activation during apoptosis (Faleiaro *et al.*, 1997), whereas ABP possessing an activated epoxide has been synthesized to follow the cathepsin activities with the progression of skin cancer (Greenbaum *et al.*, 2000). Both caspase and cathepsin are cysteine proteases, while serine proteases are profiled by the use of fluorophosphonate (Liu *et al.*, 1999).
2. *Suicide substrate probes*: This type of ABP contains a masked electrophile that is activated after the probe functions as substrate for the target enzyme. The activated electrophile then reacts with the nearby nucleophilic residue(s) in the active site such as 4-difluoromethylphenyl-*bis*(cyclohexylammounium)phosphate to probe phosphatases (Bentley *et al.*, 2002).
3. Affinity alkylating probes contains affinity based labeling groups for AFBP that require only a strong nucleophile or electrophile in the vicinity of the active cleft/crevice and do not require the enzyme to be fully active. Probes carrying this type of reactive group must rely on the selectivity of the probe scaffold to direct modification to specific enzyme/protein families. These include acyloxymethyketone for serine proteases, epoxide for cysteine proteases, and p-fluoromethylphenylglycoside for glycosidases (Campbell and Szardenings, 2003).
4. General alkylating probes contain nonspecific alkylating group that reacts with targets based only on the intrinsic reactivity of a specific amino acid such as iodoacetamido group for cysteine. This type of AFBP is valuable for tagged MS analysis of proteomes.

16.6.7.2 Linker. The linker connects the reactive group to the tag with the primary function of providing enough space between the two chemical components of an ABP/AFBP to prevent steric hindrance. This is often accomplished using a long-chain alkyl or polyethylene glycol spacer. The former is useful to modulate hydrophobicity and allow entry into cells, whereas the latter can confer solubility to hydrophobic probes in aqueous solutions. The linker can also incorporate specificity elements used to target the probe to desired enzyme/protein families. For example, a peptide or peptide type structure can be used to provide specific binding of the probe to protease active regions. The technique of preparing AFBP from recombinant proteins can be useful in the design of probes for investigating protein-binding domains that require substantial protein recognition elements for specificity.

16.6.7.3 Tag. The tag on ABP/AFBP allows quick and simple identification and purification of probe-modified proteins. The tag is the primary element that distinguishes an ABP/AFBP from a stand-alone mechanism/affinity based reagent. The most commonly used tags are affinity (biotin), fluorescent and radioactive tags. Since the simplest and most common method for protein separation involves the use of polyacrylamide gels, the tags used in probe design must be compatible with PAGE, in particular 2DE. Biotin facilitates detection by simple western blotting uses a reporter avidin. The biotin-streptavidin complex is one of the strongest known noncovalent interactions with the association constant of $10^{15}\,M^{-1}$, allowing for quantitative binding of low abundance biotinylated proteins. The fluorescent and radioactive tags can be readily visualized with high sensitivity by direct scanning of gels. Fluorescent tags also allow multiplexing of samples based on nonoverlapping excitation/emission spectra, thus permitting the use of probes with different colored tags in different experiments by obtaining all results on a single gel (Patricelli *et al.*, 2001).

The ABP/AFBP can be used on various aspects of proteomics from protein expression and identification to cellular localization and regulation. It can be applied to drug discovery in the target identification and validation (Jeffrey and Bogyo, 2003).

16.6.8 Nonsense suppression mutagenesis

Nonsense suppression is a biosynthetic approach enabling a large variety of unnatural amino acid to be incorporated site-specifically into proteins (Cornish and Schultz, 1994; Hohsaka *et al.*, 2001). The fundamental approach to nonsense (stop codon) suppression is outlined in Figure 16.10.

The site-direct mutagenesis (subsection 13.7.1) is performed by replacing the codon for the amino acid of interest with the stop codon, TAG. The subsequent transcription (section 13.4) yields UAG containing mRNA Separately, a suppressor tRNA that recognizes this codon is aminoacylated with the desired modified/unnatural amino acid. Addition of the mutagenized mRNA and the aminoacylated suppressor tRNA to an *E. coli* or rabbit reticulocyte extract capable of supporting protein biosynthesis (section 13.5) generates a mutant protein containing modified/unnatural amino acid at the desired position (Ellman *et al.*, 1992). For the *in vivo* synthesis of mutant membrane proteins, mutagenized mRNA and modified/unnatural aminoacyl tRNA are co-injected into *Xenopus* oocyte for expression (Sake *et al.*, 1996). Three issues concerning tRNA need to be considered:

1. *Orthogonality of tRNA*: The suppressor tRNA must not be recognized by aminoacyl tRNA synthetases (aRSs) of the translation system. Otherwise the suppressor tRNA, once delivering its unnatural amino acid, will be charged with a natural amino acid and return to the protein synthesis cycle. This tRNA orthogonality can be solved by the use of yeast tRNAs in the *E. coli* translation system because yeast tRNAs are not recognized by *E. coli* aRSs.

2. *Consumption of aminoacyl-tRNA*: Unnatural aminoacyl tRNA is consumed as a stoichiometric reagent in the process. Its synthesis and supply to support the mutant protein synthesis can be a tough task.

3. *Stop codon usage*: Among three stop codons, only the amber stop codon, UAG can be used for the suppression, thus multiple incorporation of unnatural amino acids into single protein cannot be expected. Though separate chains, each containing one unnatural amino acid, can be synthesized and then joined by either chemical ligation or EPL. An alternative approach exploits the extended codon-anticodon pairs, e.g. four-base codons (Hohsaka and Sisido, 2002) or non-natural base pairs that are orthogonal to the natural bases (Hirao *et al.*, 2002).

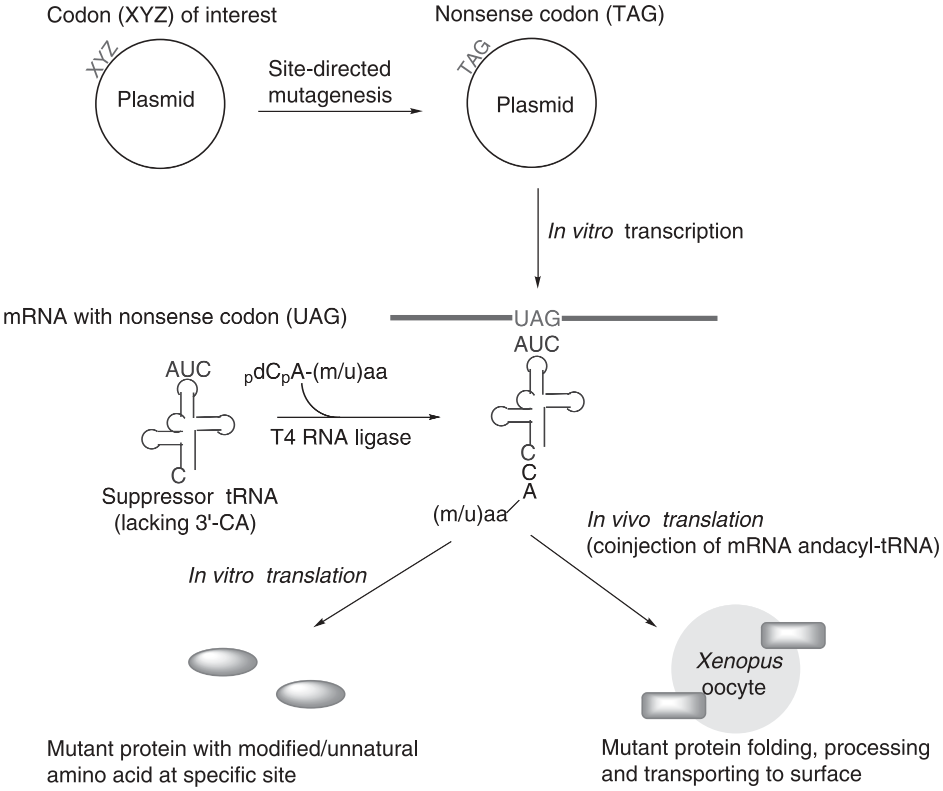

Codon (XYZ) of interest
XYZ
Plasmid
Site-directed mutagenesis
Nonsense codon (TAG)
TAG
Plasmid
In vitro transcription
mRNA with nonsense codon (UAG)
UAG
AUC
AUC
pdCpA-(m/u)aa
T4 RNA ligase
C
Suppressor tRNA (lacking 3'-CA)
C
C
A
(m/u)aa
In vivo translation (coinjection of mRNA andacyl-tRNA)
In vitro translation
Xenopus oocyte
Mutant protein with modified/unnatural amino acid at specific site
Mutant protein folding, processing and transporting to surface

Figure 16.10 Nonsense suppression mutagenesis
The nonsense suppression mutagenesis entails mutating a codon of interest to the nonsense stop codon, TAG (amber). This is followed by *in vitro* transcription of UAG-containing mRNA which directs the incorporation of modified/unnatural amino acid (m/u-amino acid) during *in vitro* or *in vivo* (by connecting mRNA and m/u-acyl tRNA) translation. The m/u-amino acid is aminoacylated onto the suppressor tRNA with anticodon, CUA which recognizes the amber codon on the mRNA. Adapted from Cornish and Schultz (1994)

A wide variety of amino acid analogs (modified/unnatural amino acids) have been introduced into proteins using nonsense suppression or protein ligation for various applications (Table 16.17). Some of potential applications of unnatural amino acids are summarized as follows:

1. Isoelectronic and isosteric amino acid analogs are extremely useful for investigating structural and functional roles of individual amino acids in proteins. For example, homoglutamate (isoelectronic) and 4-nitro-2-amino butyrate (isosteric to Glu) substitutions of Glu43 known as catalytically important in staphylococcal nuclease indicate that the essentially of Glu43 is primarily structural (Cornish and Schultz, 1994). A similar approach should be applicable to differentiate ion-pair formation versus double displacement mechanism in the retaining glycosidases.

2. Homologous amino acids with varied chain length, branch and/or cyclic structures can be used to systematically investigate the steric and polarity/hydrophobic effects of amino acids in proteins. Such studies indicate that amino acids that increase the bulk of the buried hydrophobic surface area without concomitant introduction of strain can increase protein stability significantly.

TABLE 16.17 Some amino acid analogs introduced into proteins by nonsense suppression or protein ligation and their possible applications

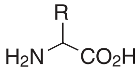
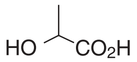
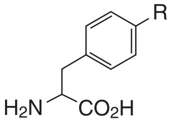
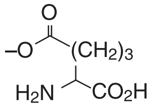
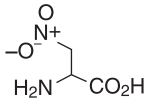
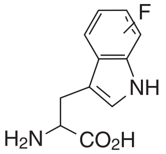
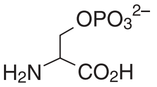
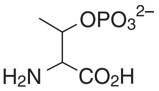
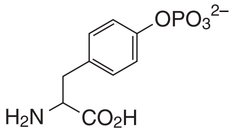
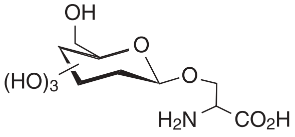
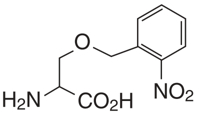
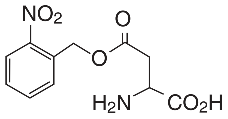
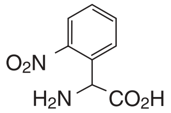
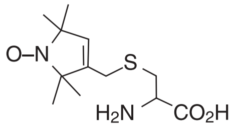
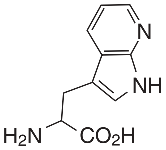
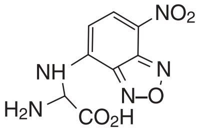
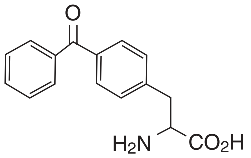
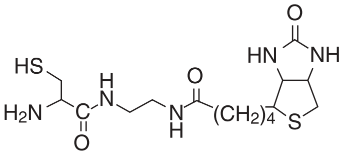
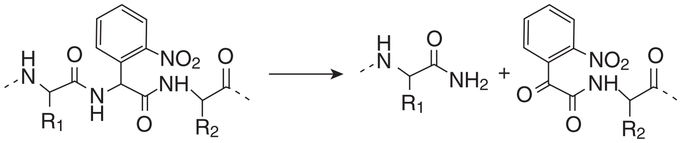
Structure of amino acid analog	Target amino acid	Applications
	Aliphatic amino acids	Alkylglycine with R groups varied in chain length and branch to probe steric and hydrophobic effects
	Gly	Lactate which is incorporated into proteins in an esteric linkage is used to probe the main chain conformation, also acts as Gly isosteric probe
	Phe	4-Substituted Phe with varied substituents to probe electronic and steric effects
	Glu/Asp	Homoglutamic acid which is isoelectonic to Glu/Asp with one/tow additional CH_2 to probe structure/functional effects of Glu/Asp
	Glu	4-Nitro-2-aminobutyric acid which is isosteric to Glu but poorer base to probe the basicity of functional Glu
	Trp	Fluorotryptophan with fluoro substituent which is isosteric to hydrogen at various/multiple positions of indole ring to probe isoelectonic and hydrophobic effects of Trp
	Ser	Phosphoserine to investigate effects of phosphorylation at Ser in PTMP and signal transduction
	Thr	Phosphothreonine to investigate effects of phosphorylation at Thr in PTMP and signal transduction
	Tyr	Phosphotyrosine to investigate effects of phosphorylation at Tyr in PTMP and signal transduction
	Ser	Glycosylserine to investigate effects of posttranslational glycosylation at Ser
	Ser	o-Nitrobenzylserine in caged proteins to probe structural and functional contributions of Ser
	Asp	β-o-Nitrobenzylaspartate in caged proteins to probe structural and functional contributions of Asp
	Gly	o-Nitrobenzylglycine in caged proteins which undergo photolytic proteolysis at Gly

TABLE 16.17 ***continued***

Structure of amino acid analog	Target amino acid	Applications
	Cys	L-2-Amino-3-thiometyl-1-(1-oxy-2,2,5,5-tetramethyl-3-pyrrolin-3-yl)-propanoic acid as a spin label in the structural and dynamic studies of proteins by electron spin resonance spectroscopy
	Trp	7-Azatryptophane as a fluorescence probe in the structural studies of proteins by fluorescence spectroscopy
	Ala	Nitrobenzoxadiazole derivative of aminoglycine as a fluorescence probe in structural studies of proteins by fluorescence spectroscopy
	Phe	p-Benzoylphenylalaine as a cross-linking agent in site-site/protein-protein interaction studies
	Cys	Cysteinyl-biotin as a tag for high efficient purification of proteins

Notes: 1. Abbreviation used: PTMP, posttranslational modification of proteins.

2. Most of the caged amino acids in caged proteins reverted to the natural amino acids upon photodecaging except o-nitrobenzylglycine which leads to the cleavage of peptide chain:



3. It is known that ribosome can mediate ester formation and nonsense suppression strategy is well-suited to this tactic. Thus α-hydroxy acids can be used to replace α-amino acids in a translational process that produces a backbone ester in place of the usual amide bond to probe protein secondary structures, because the backbone —NH-CO— that is displaced by —O-CO—, is crucial to establishing both the α-helix and the β-sheet. In addition to removing the —NH— from participating in the backbone hydrogen bonding, the amide-to-ester mutation also makes the —CO— a poorer hydrogen bond acceptor. For example, the Ala82/Lactate mutation in T4 lysozyme (Ala82 is located at a break between two helixes), causes the 4.18 kJ mol^{-1} (3.7°C) destabilization (Cornish and Schultz, 1994).

4. Modified amino acids that duplicate or mimic PTMs of proteins provide a useful means for studying *in vivo* structures, functions and regulations of proteins. Since recombinant proteins are not posttranslationally modified, introductions of the modified amino acids to duplicate posttranslationally modified proteins are invaluable to functional proteomics. They also offer avenues for exploring processes of PTM and signal transduction associated with such modifications.

5. Incorporation of amino acids with photoreactive side chains into protein is a useful application, especially involving caged amino acids in which a heteroatom is protected as an *O*-nitrobenzyl group (an inactive precursor). Illumination with light of an appropriate wavelength removes the nitrobenzyl and reveals the previously caged functionally (activation). For example, nonsense mutagenesis of Asp20 of T4 lysozyme with β-*O*-nitrobenzylaspartate results in inactivation. However, photolysis (decaging) of the mutant enzyme restores the full catalytic activity (Cornish and Schultz, 1994). The photoreactive caged amino acids can be most effectively applied to investigate the catalytic/functional residues of enzymes/proteins.

6. A wide variety of physicochemical probes can be incorporated to investigate dynamics of protein structures. Amino acid derivatives attached with spectroscopic reporters, which are sensitive to microenvironments such as ^{13}C label for NMR, spin label (e.g. N-oxy-2,2,5,5-tetramethylpyrrolinyl) for electron spin resonance and fluorophore for fluorescence spectroscopy, are commonly employed probes. Spectral changes in the probe reporters accompanying structural alternation of proteins upon molecular interactions, ligand bindings or during biochemical processes can be monitored in real time spectrophotometrically. Fluorescence energy transfer between the fluorescent ligand and the probe can be used to obtain valuable distance information relating to the binding site to the selected residue(s) of proteins.

An alternative approach to incorporate spectroscopic probes can be accomplished by the use of biarsenical tetracysteine system (Adams *et al.*, 2002) or enzymatic alkylation (Keppler *et al.*, 2003). A mutant O^6-alkylguanine-DNA alkyltansferase can efficiently transfer a benzyl group to itself when presented with O^6-benzylguanine derivatives. This enables the labeling of spectroscopic probes on to the transferase, which is fused to the target protein under investigation.

7. Amino acid analogs with a reactive functionality afford cross-linking between the selected (mutated) residue and the adjacent residues. Such study may be used to infer topologies inter- and intra-molecularly within the mutant protein or between the interacting biomacromolecules.

8. An interaction between biotin and avidin is one of the strongest biomolecular interacting systems. Therefore an introduction of biotin derivatives to proteins facilitates their efficient purification/characterization.

16.7 METABOLOME

The metabolome, which refers to the full complement of metabolites (<2 kDa) within a cell, tissue or organism (Tomita and Nishioka, 2005), comprises a diverse array of molecular chemotypes including carbohydrates, lipids, amino acids and peptides, nucleotides and derivatives, anabolic intermediates as well as catabolic products. Metabolomics aims to develop and apply strategies for the global analysis of the cellular and biofluid metabolites (Tomita and Nishioka, 2005). By integrating these small biomolecule profiles with transcript and protein expression/activity patterns, the contribution of specific metabolites and metabolic pathways to health and disease may be elucidated. An early example of such a contribution is a concept of the 'inborn error of metabolism' in which a hereditary disease is caused by a build-up of a metabolite due to a defect in the key enzyme. Metabolomics has been proposed for studies of molecular physiology, functional genomics and proteomics, clinical biochemistry, biomarker discovery, research on the mode of drug action and monitoring of drug therapy.

The global strategies to profile the metabolome must overcome a set of challenges. Unlike transcripts and proteins, metabolites share no direct link with the genetic code and are instead products of the concerted action of many networks of enzymatic reactions. Furthermore, metabolites constitute a structurally diverse collection of molecules with widely varied chemical and physical properties. Because of their chemical complexity and heterogeneity, metabolites do not readily lend themselves to universal methods for analysis and characterization. Metabolome analyses not only need to accommodate the high diversity of biomolecules but also need to cover the vast dynamic range of metabolic concentrations. Extreme care and fast inactivation of all biochemical reactions during sampling is vital to reproducible and robust metabolome analysis. Four analytical approaches for metabolomics, as given in Table 16.18, are conceivable (Birkemeyer *et al.*, 2005).

Metabolites can be analyzed by the standard tools of chemical analysis such as spectroscopic measurements (Chapter 7) and MS. The resolution, sensitivity and selectivity of these technologies can be enhanced or modified by coupling to chromatographic separations such as gas and liquid chromatography (GC and LC). For comprehensive analysis of the metabolome, fingerprinting and profiling approaches are aimed at detection of all metabolites that fall within the range of the chosen technologies. It is essential to use strategies that have a wider coverage in terms of type and number of metabolites analyzed. Prefractionation steps and subsequent parallel measurements are required to optimize analyses and to facilitate the detection of minor components or changes of metabolites. The major appeal of these approaches is the potential of discovery. Novel or unexpected metabolites can be linked to physiological processes or gene function and used as biomarkers. For targeted (quantitative) analysis of metabolite pool size and flux, the structure identification and quantification of concentrations are considered/preconceived because only information on the metabolites under scrutiny is retrieved. The development of MS capable of resolving mass isotopomers (each mass variants of a chemical substance, e.g. single unit mass differences of isotopes) greatly facilitate the use of stable isotopes in investigations of metabolite flux (Birkemeyer *et al.*, 2005).

The metabolite profiling has primarily been performed with NMR and/or MS techniques used alone or in combination with GC/LC. NMR offers a rapid and noninvasive method to comparatively characterize metabolite expression pattern (Shulman and Rothman, 2005). However, NMR shows limited sensitivity and resolution and is best suited for detecting the most abundant metabolites in complex sample. Similar limitations are also encountered with MS (Allen *et al.*, 2003; Fiehn *et al.*, 2000). The sensitivity and metabolome coverage by MS analysis can be greatly increased by using GC-MS and LC-MS experiments. Each of the profiling methods can be conducted in either a targeted or untargeted mode. Targeted applications focus on the characterization of a specific class of metabolites by exploiting their unique chemical properties. By contrast, untargeted applications seek to broadly profile the metabolome by establishing conditions for the concurrent analysis of as many metabolites as possible. Both are complementary in interrogating the metabolome and assess the biochemical repercussions of specific genomic, proteomic and/or physiological perturbations.

In its relationship to functional proteomics, the global metabolite profiling offers a potentially powerful strategy to the functional assignment of enzyme networks because:

- Many enzymes function as parts of large protein complexes and networks *in vivo*, which may be difficult to reconstruct *in vitro*.

TABLE 16.18 Overview of general approaches to metabolome analysis

	Fingerprinting	Profiling	Pool size analysis	Flux analysis
Type of analysis	Comprehensive	Comprehensive	Targeted	Targeted
Field of application	Functional gen-/prot-omics, diagnostics	Functional gen-/prot-omics, mol. physiol.	Biochem, biotechnol, mol. physiol.	Biochem, biotechnol, modeling
Major result	Metabolite classification	Relative quantif of change in metabolite pool, discovery of biomarkers, novel metabolites	Absolute quantif of metabolite pools	Quantif of metabolite flux
Metabolite covered	Limited by choice of extraction protocol, analytical method	Limited by choice of extraction protocol, analytical method	Predefined set of targeted metabolites	Predefined set of targeted metabolites
Metabolite identification	Metabolite identif not required	Identif of as many metabolites as possible	Unambiguous metabolite identif required	Unambiguous metabolite identif required
Metabolite concentration	The conc of the most abundant metabolite determines the upper sample load. The dynamic range of the instrument defines the detection limit of the minor metabolites	The conc of the most abundant metabolite determines the upper sample load. The dynamic range of the instrument defines the detection limit of the minor metabolites	Prepurification enables analysis of trace metabolites and choice of the sensitivity of the analytical instrument	Prepurification enables analysis of trace metabolites and choice of the sensitivity of the analytical instrument
Sample preparation	Mixture or pre-fractionation	Mixture or pre-fractionation	Pre-treatment incl. selective purification	Pre-treatment incl. selective purification
Analytical method	NMR or MS or CM	NMR or MS or CM	Combination of GC/LC with MS and/or NMR	Combination of GC/LC with MS and/or NMR
Sample throughput	High	High-medium	Varied	Medium-low
Reference	a, b, c	c, d, e	b, f	g, h, i

Notes: 1. Abbreviations used: mol. physiol molecular physiology, quantif, quantification; CM, chromatographic method(s);
2. References are: [a] Oliver *et al*. (1998); [b] Fiehn (2002); [c] Goodacre *et al*. (2004); [d] Kopka *et al*. (2004); [e] Ilyin *et al*. (2004); [f] Roessner *et al*. (2001); [g] Wiechert (1998); [h] Wittmann (2002); [i] Sauer (2004).

- Enzymes are often regulated by posttranslational modifications *in vivo* which may alter substrate recognition and catalysis.
- The *in vivo* substrates and effectors of enzymes may differ from those of the *in vitro*. The novel natural substrates/effectors/products can only be detected/analyzed by the global metabolite profiling.

Thus metabolomics provides access to a portion of biomolecular space (the metabolome) that is inaccessible to genomics and proteomics, enabling the assignment of endogenous biochemical functions to a broad range of enzymes (Saghatelian and Cravatt, 2005).

16.8 REFERENCES

Adams, S.R., Campbell, R.E., Gross, L.A. *et al.* (2002) *Journal of the American Chemistry Society*, **124**, 6063–76.

Aebersold, R. and Goodlett, D.R. (2001) Mass spectrometry in proteomics. *Chemistry Reviews*, **10**, 269–5.

Aitken, A. (1990) *Identification of Protein Consensus Sequences*, Ellis Horwood Ltd. New York.

Alexander, M.K., Bourns, B.D. and Zakian, V.A. (2001) *Methods in Molecular Biology*, **177**, 241–59.

Allen, J., Davey, H.M., Broadhurst, D. *et al.* (2003) *Nature Biotechnology*, **21**, 692–6.

Anand, G.S., Law, D., Mandell, J.G. *et al.* (2003) *Proceedings of the National Academy of Sciences, USA*, **100**, 13264–9.

Andreeva, A., Howorth, D., Brenner, S.E. *et al.* (2004) *Nucleic Acid Research*, **32**, D226–9.

Appel, R.D., Bairoch, A. and Hochstrasser, D.F. (1994) *Trends in Biochemical Science*, **19**, 258–60.

Appel, R.D., Bairoch, A., Sanchez, J.-C. *et al.* (1994) *Proteins*, **18**, 19–33.

Attwood, T.K., Beck, M.E., Flower, D.R. *et al.* (1998) *Nucleic Acids Research*, **26**, 304–8.

Augen, J. (2004) *Bioinformatics in the Post-Genome Era: Genome, Transcriptome, Proteome and Information-based Medicine*, Addison Wesley Professional, Boston, MA.

Bairoch, A. and Apweiller, R. (2000) *Nucleic Acids Research*, **28**, 45–8.

Bairoch, A., Apweiller, R., Wu, C.H. *et al.* (2005) *Nucleic Acids Research*, **33**, D154–9.

Barker, W.C., Garawelli, J.S., Hou, Z. *et al.* (2001) *Nucleic Acids Research*, **29**, 29–32.

Barry, R. and Soloviev, M. (2004) *Proteomics*, **4**, 3717–26.

Bateman, A., Birney, E., Durbin, R. *et al.* (2000) *Nucleic Acids Research*, **28**, 263–6.

Beavis, R.C. and Chait, B.T. (1996) *Methods in Enzymology*, **270**, 519–51.

Bentley, J.R., Cesaro-Tadic, S., Mekhalfia, A. *et al.* (2002) *Angew. Chem. Int. Ed. Engl.*, **41**, 775–7.

Bergfors, T.M. (ed.) (1999) *Protein Crystallization*, International University Line, La Jolla, CA.

Berman, H., Westbrook, J., Feng, Z. *et al.* (2002) *Nucleic Acids Research*, **28**, 235–42.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B. *et al.* (1977) *Journal of Molecular Biology*, **112**, 535–42.

Birkemeyer, C., Luedemann, A., Wagner, C. *et al.* (2005) *Trends in Biotechnology*, **23**, 28–33.

Blackstock, W.P. and Weir, M.P. (1999) *Trends in Biotechnology*, **17**, 121–7.

Bleasby, A.J., Akrigo, D. and Attwood, T.K. (1994) *Nucleic Acids Research*, **22**, 3573–7.

Bloomberg, A., Bloomberg, L., Norbeck, J. *et al.* (1995) *Electrophoresis*, **16**, 1935–45.

Blundell, T.L., Jhoti, H. and Abell, C. (2002) *Nature Reviews in Drug Discovery*, **1**, 45–54.

Bonneau, R. and Baker, D. (2001) *Annual Reviews in Biophysical Biomolecular Structure*, **30**, 173–89.

Bovie, J., Lüthy, R. and Eisenberg, D. (1991) *Science*, **253**, 164–70.

Brändén, C.-I. (1980) *Quarterly Reviews in Biophysics*, **13**, 317–38.

Brenner, S.E., Koehl, P. and Levitt, M. (2000) *Nucleic Acids Research*, **28**, 254–6.

Brooks, C.I. III, Karplus, M. and Pettitt, B.M. (1988) *Proteins: A Theoretical Prospective of Dynamics, Structure, and Thermodynamics*, John Wiley & Sons, New York.

Brown, R.S. and Lennon, J.J. (1995) *Analytical Chemistry*, **67**, 1998–2003.

Brown, N.P., Orengo, C.A. and Taylor, W.R. (1996) *Computer Chemistry*, **3**, 359–80.

Bryant, S.H. and Altschul, S.F. (1995) *Cun. Opin. Struct. Biol.* **5**, 236–44.

Campbell, D.A. and Szardenings, A.K. (2003) *Current Opinions in Chemistry and Biology*, **7**, 296–303.

Castrignano, T., Demeo, P.D., Cozetto, D., Talamo, I.G. and Tramontano, A. (2006). *Nucleic Acid Resear.* **34**, D306–9.

Chinea, G., Padron, G., Hooft, R.R.W. *et al.* (1995) *Proteins*, **23**, 415–28.

Chothia, C. and Lesk, A.M. (1986) *EMBO Journal*, **5**, 823–6.

Claude, J.-B., Suhre, K., Notredame, C. *et al.* (2004) *Nucleic Acids Research*, **32**, W606–9.

Combet, C., Blanchet, C., Geourjon, C. and Deléage, G. (2000) *Trends in Biochemical Science*, **25**, 147–50.

Comeau, S.R., Gatchell, D.W., Vajda, S. and Camacho, C.J. (2004) *Nucleic Acids Research*, **32**, W96–9.

Cornish, T.J. and Cotter, R.J. (1997) *Analytical Chemistry*, **69**, 4615–8.

Cornish, V.W. and Schultz, P.G. (1994) *Current Opinions in Structural Biology*, **4**, 601–7.

Cravatt, B.F. and Sorensen, E.J. (2000) *Current Opinions in Chemical Biology* **4**, 663–8.

Cunningham, B.C., Jhurani, P., Ng, P. and Wells, J.A. (1989) *Science*, **243**, 1330–6.

Dandekar, T. and Argos, P. (1994) *Journal of Molecular Biology*, **236**, 844–61.

Deshpande, N., Addess, K.J., Bluhm, W.F. *et al.* (2005) *Nucleic Acids Research*, **33**, D233–7.

Ellman, J.A., Mendel, D., Anthony-Cahil, S. *et al.* (1992) *Methods in Enzymology*, **202**, 301–36.

Enright, A.J., Iliopoulos, I., Kyrpides, N.C. and Ouzounis, C.A. (1999) *Nature*, **402**, 86–90.

Faleiro, L., Kobayashi, R., Fearmhead, H. and Lazebnik, Y. (1997) *EMBO Journal*, **16**, 2271–81.

Feig, M. and Brooks, C.L. III (2004) *Current Opinions in Structural Biology*, **14**, 217–24.

Fenn, J.B., Mann, M., Meng, C.K. *et al.* (1989) *Science*, **246**, 64–71.

Ficarro, S.B., McCleland, M.L., Stukenberg, P.T. *et al.* (2002) *Nature Biotechnology*, **20**, 301–5.

Fiehn, O., Kopka, J., Dormann, P. *et al.* (2000) *Nature Biotechnology*, **18**, 1157–61.

Fiehn, O. (2002) *Plant Molecular Biology*, **48**, 155–71.

Fung, E. (ed.) (2004) *Protein Arrays: Methods and Protocols*, Humana Press, Totowa, NJ.

Gasteiger, E., Gattiker, A., Hoogland, C.I. *et al.* (2003) *Nucleic Acids Research*, **31**, 3784–8.

Gatti, A. and Traugh, J.A. (1999) *Analytical Biochemistry*, **266**, 198–204.

Goodacre, R., Vaidyanathan, S., Dunn, W.B. *et al.* (2004) *Trends in Biotechnology*, **22**, 245–52.

Görg, A., Weiss, W. and Dunn, M.J. (2004) *Proteomics*, **4**, 3665–85.

Gray, J.J., Moughon, S.E., Kortemme, T. *et al.* (2003) *Proteins*, **52**, 118–22.

Greenbaum, D., Medzihradszky, K.F., Burlingame, A. and Bogyo, M. (2000) *Chemical Biology*, **7**, 569–81.

Greer, J. (1990) *Science*, **214**, 149–59.

Griffin, J.L. and Shockcor, J.P. (2004) *Nature Rev Cancer*, **4**, 551–6.

Gronborg, M., Kristiansen, T.Z., Stensballe, A. *et al.* (2002) *Molecular Cell Proteomics*, **1**, 517–27.

Guex, N., Diemand, A. and Peitsch, M.C. (1999) *Trends in Biochemical Science*, **24**, 364–7.

Guo, J.-T., Ellrott, K., Chung, W.J. *et al.* (2004) *Nucleic Acids Research*, **32**, W522–5.

Gupta, R., Birch, H., Rapacki, K. *et al.* (1999) *Nucleic Acids Research*, **27**, 370–2.

Gygi, S.P., Rochon, Y., Franz, B.R. and Aebersold, R. (1999) *Molecular Cell Biology*, **19**, 1720–30.

Halligan, B.D., Ruotti, V., Jin, W.S. *et al.* (2004) *Nucleic Acids Research*, **32**, W638–44.

Hamdan, M. and Righetti, P.G. (2005) *Proteomic Today: Protein Assessment and Biomarkers Using Mass Spectrometry, 2D Electrophoresis and Microarray Technology*, John Wiley & Sons, Hoboken, NJ.

Haustein, E. and Schwille, P. (2004) *Cun. Opin. Struct. Biol.* **14**, 531–40.

Hengen, P.N. (1997) *Trends in Biochemical Science*, **22**, 33–4.

Henikoff, S., Pietrokowski, S. and Henikoff, J.G. (1998) *Nucleic Acids Research*, **26**, 309–12.

Herbert, B. (1999) *Electrophoresis*, **20**, 660–3.

Higgins, D. and Taylor, W. (eds) (2000) *Sequence, Structure and Databanks: A Practical Approach*, Oxford University Press, Oxford, UK.

Hilbert, M., Bohm, G. and Jaenicke, R. (1993) *Proteins*, **17**, 138–51.

Hirao, I., Ohtsuki, T., Fujiwara, T. *et al.* (2002) *Nature Biotechnology*, **20**, 177–82.

Hobohm, U. and Sander, C. (1995) *Journal of Molecular Biology*, **251**, 390–9.

Hofmann, K., Bucher, P., Falquet, L. and Bairoch, A. (1999) *Nucleic Acids Research*, **27**, 215–9.

Hohsaka, T., Ashizuka, Y., Taira, H. *et al.* (2001) *Biochemistry*, **40**, 11060–4.

Hohsaka, T. and Sisido, M. (2002) *Cun. Opin. Chem. Biol.* **6**, 809–15.

Hulo, N., Bairoch, A., Bulliard, V. *et al.* (2006) *Nucleic Acids Research*, **34**, D227–30.

Hultschig, C., Kreutzberger, J., Seitz, H., Konthm, Z., Büssow, K. and Lehrach, H. (2006) *Cun. Opin. Chem. Biol.* **10**, 4–10.

Hung, L.-H. and Samudrala, R. (2003) *Nucleic Acids Research*, **31**, 3296–9.

Ilyin, *et al.* (2004) *Trends in Biotechnology*, **22**, 411–6.

Ito, M., Matsuo, Y. and Nishikawa, K. (1997) *Computer Applied Bioscience* (*CABIOS*), **13**, 415–23.

Jeffery, D.A. and Bogyo, M. (2003) *Current Opinions in Biotechnology*, **14**, 87–95.

Jensen, O.N. (2004) *Current Opinions in Chemical Biology*, **8**, 33–41.

Jensen, O.N., Podtelejnikow, A.V. and Mann, M. (1997) *Analytical Chemistry*, **69**, 47–50.

Johnson, M.S., May, A.C.W., Rodinov, M.A. and Overington, J.P. (1996) *Methods in Enzymology*, **266**, 575–98.

Jones, G., Willett, P., Glen, R.C., Leach, A.R. and Taylor, R. (1997) *J. Mol. Biol.* **267**, 727–48.

Jona, G. and Synder, M. (2003) *Current Opinions in Molecular Therapy*, **5**, 271–7.

Jones, D., Taylor, W. and Thornton, J. (1992) *Nature*, **258**, 86–9.

Jorgensen, W.L., Chandrasekhas, J., Madura, J.D. *et al.* (1983) *Journal of Chemical Physics*, **79**, 926–35.

Kaufmann, R. (1995) *Journal of Biotechnology*, **41**, 155–75.

Kawashima, S. and Kanehisa, M. (2000) *Nucleic Acids Research*, **28**, 374.

Kelly, L.A., MacCallum, R.M. and Sternberg, M.J.E. (2000) *Journal of Molecular Biology*, **299**, 501–22.

Keppler, A., Gandreizig, S., Gronemyer, T. *et al.* (2003) *Nature Biotechnology*, **21**, 86–9.

Kersey, P., Bower, L., Morris, L. *et al.* (2005) *Nucleic Acids Research*, **33**, D297–302.

Kinter, M. and Sherman, N.E. (2000) *Protein Sequencing and Identification Using Tandem Mass Spectrometry*, John Wiley & Sons, Inc. New York.

Klose, J. (1999) *Methods in Molecular Biology*, **112**, 147–72.

Kopka, J. *et al.* (2004) *Genome Biology*, **5**, 109–17.

Koppensteiner, W.A., Lackner, P., Wiederstein, M. and Sippl, M.J. (2000) *Journal of Molecular Biology*, **296**, 1139–52.

Kouranov, A., Xie, L., deLacurz, J., Chen, L., Westbrook, J., Bourne, P.E. and Berman, H.M. (2006) *Nucleic Acid Resear*, **34**, D302–5.

Koza, J. (1993) Genetic Programming, MIT Press, Cambridge, MASS.

Krebs, E.G. and Beavo, J.A. (1979) *Annual Reviews in Biochemistry*, **48**, 923–59.

Kreegipuu, A., Blom, N. and Brunak, S. (1999) *Nucleic Acids Research*, **27**, 237–9.

Kurowski, M.A. and Bujnicki, J.M. (2003) *Nucleic Acids Research*, **31**, 3305–7.

LaBaer, J. and Ramachandran, N. (2005) *Current Opinions in Chemical Biology*, **9**, 14–9.

Laskowski, R.A. (2001) *Nucleic Acids Research*, **21**, 221–2.

Laskowski, R.A., Chistyakow, V.V. and Thornton, J.M. (2005) *Nucleic Acids Research*, **33**, D266–8.

Lecrenier, N., Foury, F. and Goffeau, A. (1998) *BioEssays*, **20**, 1–6.

Lee, J., Liwo, A., Ripoll, D. *et al.* (1999) *Proteins*, **S3**, 177–85.

Lee, Y.S. and Mrksich, M. (2002) *Trends in Biotechnology*, **20**, 514–8.

Lesk, A.M. and Chothia, C. (1984) *Journal of Molecular Biology*, **174**, 175–91.

Letunic, I., Copley, R.R., Pils, B. *et al.* (2006) *Nucleic Acids Research*, **34**, D257–60.

Levine, M., Stuart, D. and Williams, J. (1984) *Acta Crystallography*, **A40**, 600–60.

Levitt, M. and Chothia, C. (1976) *Nature*, **261**, 552–8.

Licitra, E.J. and Liu, J.O. (1996) *Proceedings of the National Academy of Sciences, USA*, **93**, 12817–21.

Liebler, D.C. (2002) *Introduction to Proteomics*, Humana Press, Totowa, NJ.

Liu, Y., Patricelli, M.P. and Cravatt, B.F. (1999) *Proceedings of the National Academy of Sciences, USA*, **96**, 14694–9.

Lundblad, R.L. (2005) *Proteomics: Applications in Solution Protein Chemistry*, Taylor & Francis, Boca Raton, FL.

Mann, M., Hendrickson, R.C. and Pandey, A. (2001) *Annual Reviews in Biochemistry*, **70**, 437–73.

May, A.C. (1999) *Proteins*, **37**, 20–9.

McDermott, J. and Samudrala, R. (2003) *Nucleic Acids Research*, **31**, 3736–7.

Mendez, R., Leplae, R., DeMaria, L. and Wodak, S.J. (2003) *Proteins*, **52**, 51–67.

Miller, J.C., Zhou, H., Kwekel, J. *et al.* (2003) *Proteomics*, **3**, 56–63.

Milpetz, F., Argos, P. and Persson, B. (1995) *Trends in Biochemistry Science*, **20**, 204–5.

Mitchell, P. (2002) *Nature Biotechnology*, **20**, 225–9.

Moult, J., Fidelis, K., Zemla, A. and Hubbard, T. (2003) *Proteins*, **53**(suppl. 6), 334–9.

Murzin, A. and Bateman, A. (1997) *Proteins*, **29S**, 105–12.

Needleman, S.B. and Wunsch, C.D. (1970) *J. Mol. Biol.* **48**, 443–53.

Nielsen, H., Brunak, S. and vonHeijne, G. (1999) *Protein Eng.* **12**, 3–9.

Oliver, S.G. *et al.* (1998) *Trends in Biotechnology*, **16**, 373–8.

Ortiz, A., Kolinkski, A., Rotkiewicz, P. *et al.* (1999) *Proteins*, **S3**, 177–85.

Patricelli, M.P., Giang, D.K., Stamp, L.M. and Burbaum, J.J. (2001) *Proteomics*, **1**, 1067–71.

Pearl, F., Todd, A., Sillitoe, I. *et al.* (2005) *Nucleic Acid Research*, **33**, D247–51.

Pechkova, E. and Nocolini, C. (2003) *Proteomics and Nanocrystallography*, Kluwer Academic/Plenum Publishers, New York.

Pellegrini, M., Marcotte, E.M., Thompson, M.J. *et al.* (1999) *Proceedings of the National Academy of Sciences, USA*, **96**, 4285–8.

Peng, J., Schwartz, D., Elias, J.E. *et al.* (2003) *Nature Biotechnology*, **21**, 921–6.

Pennington, S.R. and Dunn, M.J. (eds) (2001) *Proteomics from Protein Sequence to Function*. BIOS Scientific Publishers Ltd., Springer-Verlag, New York.

Perriére, G., Combet, C., Penel, S. *et al.* (2003) *Nucleic Acids Research*, **31**, 3393–9.

Phizichy, E.M. and Field, S. (1995) *Microbiology Reviews*, **59**, 94–123.

Piehler, J. (2005) *Current Opinions in Structural Biology*, **15**, 4–14.

Porter, C.T., Bartlett, G.J. and Thornton, J.M. (2004) *Nucleic Acids Research*, **32**, D129–33.

Predki, P.F. (2004) Functional protein microarray. *Current Opinions in Chemical Biology*, **8**, 8–13.

Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2005) *Nucleic Acids Research*, **33**, D501–4.

Putz, U., Skehel, P. and Kuhl, D. (1996) *Nucleic Acids Research*, **24**, 4838–40.

Rabilloud, Th. Ed. (2000) *Proteome Research: Two-Dimensional Gel Electrophoresis and Identification Methods*, Springer, New York.

Ramachandran, N., Hainsworth, E., Bhullar, B. *et al.* (2004) *Science*, **305**, 86–90.

Ranish, J.A., Yi, E.C., Leslie, D.M. *et al.* (2003) *Nature Genetics*, **33**, 349–55.

Reinders, J., Lewandrowski, U., Moeblus, J. *et al.* (2004) *Proteomics*, **4**, 3686–703.

Rodgers, S., Wells, R. and Rechsteiner, M. (1986) *Science*, **234**, 364–8.

Roessner, U. *et al.* (2001) *Plant Cell*, **13**, 11–29.

Rossmann, M.G. and Argos, P. (1975) *Journal of Biological Chemistry*, **250**, 7525.

Rost, B. and Sander, C. (1993) *Journal of Molecular Biology*, **232**, 584–99.

Rost, B. and Liu, J. (2004) *Nucleic Acids Research*, **32**, W321–6.

Saghatelian, A. and Cravatt, B.F. (2005) *Current Opinions in Chemical Biology*, **9**, 62–8.

Sake, M.E., Sampson, J.R., Nowak, M.W. *et al.* (1996) *Journal of Biological Chemistry*, **271**, 1057–63.

Šali, A. and Blundell, T.L. (1993) *Journal of Molecular Biology*, **234**, 779–815.

Salwinski, L. and Eisenberg, D. (2003) *Cun. Opin. Struct. Biol.* **13**, 377–82.

Samudrala, R., Xia, Y., Huang, E. and Levitt, M. (1999) *Proteins*, **S3**, 194–8.

Sánchez, R., Pieper, U., Mirkovic deBakker, P.I.W. *et al.* (2000) *Nucleic Acids Research*, **28**, 250–3.

Sauer, U. (2004) *Current Opinions in Biotechnology*, **15**, 58–63.

Schuler, G.D., Epstein, J.A., Ohkana, H. and Kans. J.A. (1996) *Meth. Enzymol.* **266**, 141–62.

Schwede, T., Kopp, J., Guex, N. and Peitsch, M.C. (2003) *Nucleic Acids Research*, **31**, 3381–5.

Sheu, S.-H., Lancia, D.R. Jr., Clodfelter, K.H. *et al.* (2005) *Nucleic Acids Research*, **33**, D206–11.

Shi, J., Blundell, T.L. and Mizuguchi, K. (2001) *Journal of Molecular Biology*, **310**, 243–57.

Shindyalov, I.N. and Bourne, P.E. (1997) *Computer Applied Bioscience* (*CABIOS*), **13**, 487–96.

Shulman, R.G. and Rothman, D.L. (2005) *Metabolomics by* in vivo *NMR*, John Wiley & Sons, Hoboken, NJ.

Smith, G.R. and Sternberg, M.J. (2002) *Current Opinions in Structural Biology*, **12**, 28–35.

Sreekumar, A., Nyati, M.K., Varambally, S. *et al.* (2001) *Cancer Research*, **61**, 7585–93.

Stagljar, I. and Field, S. (2002) *Trends in Biochemical Science*, **27**, 559–63.

Stagljar, I., Korostensky, C., Johnsson, N. and Heesen, S. (1998) *Proceedings of the National Academy of Sciences, USA*, **95**, 5187–92.

Stewart, D.E., Weiner, P.K. and Wampler, J.E. (1987) *J. Mol. Graph.* **5**, 133–40.

Swindells, M.B., Orengo, C.A., Jones, D.T. *et al.* (1998) *BioEssays*, **20**, 884–91.

Tomita, M. and Nishioka, T. (eds) (2005) *Metabolomics: The Frontier of Systems Biology*, Springer, New York.

Tonge, R., Shaw, J., Middleton, B. *et al.* (2001) *Proteomics*, **1**, 377–96.

Trester-Zedlitz, M., Kamada, K., Burley, S.K. *et al.* (2003) *Journal of the American Chemistry Society*, **125**, 2416–25.

Tsai, C.S. (2001) *Journal of Chemistry Education*, **78**, 837–9.

Tsai, C.S. (2002) *An Introduction to Computational Biochemistry*, John Wiley & Sons, New York.

Tsigelny, M.I.F. (2002) *Protein Structure Prediction: Bioinformatic approach*, International University Line, LaJolla, CA.

Uetz, P., Giot, L., Cagney, G. *et al.* (2000) *Nature*, **403**, 623–7.

Waksman, G. (ed.) (2005) *Proteomics and Protein-Protein Interactions: Biology, Chemistry, Bioinformatics and Drug Design*, Springer, New York.

Wang, Y., Anderson, J.B., Chen, J. *et al.* (2002) *Nucleic Acids Research*, **30**, 249–52.

Webster, D.M. (ed.) (2000) *Protein Structure Prediction: Methods and protocols*, Humana Press, Totowa, New Jersey.

Wery, J.P. and Schevitz, R.W. (1997) *Current Opinions in Chemical Biology*, **1**, 365–9.

Westermeier, R. and Naven, T. (2002) *Proteomics in Practice: A Laboratory Manual of Proteome Analysis*, John Wiley & Sons, New York.

Wiechert, T. (1998) *Metab. Eng.*, **3**, 195–206.

Wittke, S., Dunnwald, M., Albertsen, M. and Johnsson, N. (2002) *Molecular Biology Cell*, **13**, 2223–32.

Wittmann, C. (2002) *Advances in Biochemical Engineering Biotechnology*, **74**, 39–64.

Wu, C.H., Huant, H., Arminski, L. *et al.* (2002) *Nucleic Acids Research*, **30**, 27–30.

Zhang, H., Yan, W. and Aebersold, R. (2004) *Current Opinions in Chemical Biology*, **8**, 66–75.

Zhu, H., Bilgin, M. and Snyder, M. (2003) *Annual Reviews in Biochemistry*, **72**, 783–812.

World Wide Webs cited

2DE metadatabase:	http://www-lmmb.ncifcrf.gov/flicker
3D-PSSM:	http://www.bmm.icnet.uk/~3dpssm/
AAindex database:	http://www.genome.ad.jp/aaindex/
AatDB:	http://genome-www.stanford.edu/
ASTRAL:	http://astral.stanford.edu/
aBi:	http://www.up.univ-mrs.fr/~wabim/d_abim/compo-p.html
BCM (Baylor College of Medicine):	http://www.hgsc.bcm.tmc.edu/SearchLauncher/
Bioverse:	http://bioverse.compbio.washington.edu/
BLOCKS:	http://www.blocks.fhcrc.org/
BMCD:	http://wwwbmcd.nist.gov:8080/bmcd/bmcd.html
Catalytic Site Atlas (CSA):	http://www.ebi.ac.uk/thornton-srv/databases/CSA
CaspR:	http://igs-server.cnrs-mrs.fr/Caspr/
CATH:	http://www.biochem.ucl.ac.uk/bsm/cath/
Center for Biol. Sequence Analysis (CBS):	http://www.cbs.dtu.dk/
ClusPro:	http://nrc.bu.edu/cluster/
DDBJ:	http://srs.ddbj.nig.ac.jp/index-e.html
Entrez:	http://www.ncbi.nlm.nih.gov/Entrez
Enzyme Structure Database:	http://www.biochem.ucl.ac.uk/bsm/enzymes/index.html
Expert Protein Analysis System (ExPASy):	http://www.expasy.org
ExPASy Proteomic tools:	http://www.expasy.org/tools/
FlyBase:	http://flybase.bio.indiana.edu/
FSSP:	http://www.bioinfo.biocenter.helsinki.fi:8080/dali/
FUGUE:	http://www-cryst.bioc.cam.ac.uk/~fugue/prfsearch.html
G to P server:	http://spock.genes.nig.ac.jp/~genone/gtop.html
GeneSilico:	http://genesilico.pl/meta
iMolTalk:	http://i.moltalk.org
Integr8:	http://ebi.ac.uk/integr8
InterPro:	http://www.uniprot.org/interpro
International Protein Index (IPI)	ftp://ftp.ebi.ac.uk/pub/databases/IPI/
Interaction thermodynamic data:	http://gibk26.bse.kyutech.ac.jp/jouhou/pronit/pronit.html
JOY:	http://www-cryst.bioc.cam.ac.uk/~joy/
Jpred:	http://jura.ebi.ac.uk:8888/index.html
LIBRA I:	http://www.ddbj.nig.ac.jp/E-mail/libra/LIBRA_I.html
MIPS:	http://www.mips.biochem.mpg.de/

MMBD: http://www.ncbi.nlm.nih.gov/Entrez/structure.html
ModBase: http://guitar.rockefeller.edu/modbase/
nnPredict: http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html
NPS@: http://npsa-pbil.ibcp.fr
PANAL: http://mgd.ahc.umn.edu/panal/run_panal.html
PBIL: http://pbil.univ-lyon1.fr
PDB (Protein Data Bank): http://www.rcbs.org/pdb/
PDBsum: http://www.ebi.ac.uk/thornton-srv/databases/pdbsum
PEST: http://www.icnet.uk/LRITu/projects/pest/
Pfam: http://www.sanger.ac.uk/Pfam/
PIR (Protein Information Resource)-PSD: http://pir.georgetown.edu/
PMDB: http://www.cospur.it/PMDB
PPSearch of EBI: http://www2.ebi.ac.uk/ppsearch/
PRECISE: http://precise.bu.edu/
PredictProtein: http://www.predictprotein.org
PRINTS: http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/
ProMoST: http://proteomics.mcw.edu/promost
PROPSEARCH: http://www.embl-heidelberg.de/prs.html
ProfileScan server: http://www.isrec.isb-sib.ch/software/PFSCAN_form.html
PROSITE: http://expasy.hcuge.ch/sprot/prosite.html, http://www.isrec.isb-sib.ch/
PROSPECT-PSPP: http://csbl.bmb.uga.edu/protein_pipeline
ProTherm: http://gibk26.bse.kyutech.ac.jp/jouhou/protherm/protherm.html
PROTINFO: http://protinfo.compbio.washington.edu/
PSA: http://bmerc-www.bu.edu/psa/index.html
RefSeq: http://www.ncbi.nlm.nih.gov/RefSeq
Relibase: http://relibase.ebi.ac.uk/reli-cgi/rll?/reli-cgi/query/form_home.pl
SA-Search: http://bioserv.rpbs.jussieu.fr/cgi-bin/SA-Search
SCOP: http://scop.mrc-lmb.cam.ac.uk/scop/
SMART: http://smart.embl-heidelberg.de/
SPEMZYME: http://www.expasy.org/enzyme/
SSThread: http://www.ddbj.nig.ac.jp/E-mail/ssthread/www_service.html
STING Mellennium Suite (SMS): http://trantor.bioc.columbia.edu/SMS
SWISS-2DPAGE: http://www.expasy.org/ch2d/
SWISS-MODEL: http://www.expasy.org/swissmod/SWISS-MODEL.html
SWISS-MODEL Repository: http://expasy.org/swissmod/smrep.html
Swiss-Prot (at EBI): http://www.ebi.ac.uk/
Swiss-Prot (at ExPASy): http://www.expasy.org/sprot
SWISS-PROT knowledgebase: http://www.expasy.org/sprot/
TargetP: http://www.cbs.dtu.dk/services/TargetP/
TMAP: http://www.embl-heidelberg.de/tmap/tmap_info.html
Tmpred: http://www.ch.embnet.org/software/TMPRED_form.html
TrEMBL: http://www.expasy.org/sprot
UniProt http://www.uniprot.org/
WORLD-2DPAGE index: http://www.expasy.ch/ch2d/2d-index.html
WPDB: http://www.sdsc.edu/pb/wpdb
Yeast Proteome Database (YPD): http://www.proteome.com/
General proteome resources: Table 16.1
Secondary databases for protein sequences: Table 16.3
Enzyme databases Table 16.4
Receptor and signal protein databases: Table 16.5
Protein structural databases: Table 16.6
Integrated genome/proteome databases: Table 16.8
Databases for functional and PTM sites: Table 16.9
Servers for protein secondary structure predictions: Table 16.10
Servers for protein structure alignment, similarity/overlap: Table 16.11
Servers for protein structure predictions: Table 16.12
Servers for protein specific structure features: Table 16.13
Protein interaction databases and servers: Table 16.14
MS-based protein identification tools: Table 16.15
Phage display antibody library: Table 16.16

CHAPTER 17

GLYCOMICS

17.1 FEATURES OF GLYCOMICS

17.1.1 Glycobiology: Nomenclature and representation of glycans

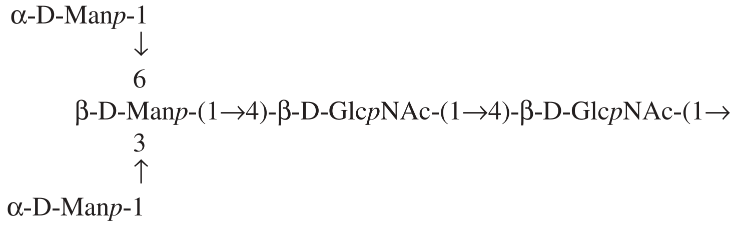
Glycan structures are complex and variable due to differences in:

- types of monosaccharide (glycose) units and modifications present;
- types of linkages; and
- the presence of branching.

Thus topologies for complex glycans differ significantly from the simple linear form describing DNA/RNA and proteins. Such structure complexity and variability create difficulties in the development of simple and consistent nomenclature and representation for glycans (for simplicity, glycans refer to both oligo- and polysaccharides). The availability of a generally accepted linear description of glycans that can be processed by computers is urgently needed in the bioinformatics era for the efficient carbohydrate data collections and cross-linking. Some progress has been made in this direction:

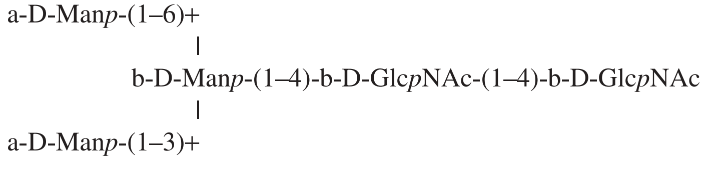
1. IUPAC-IUBMB (http://www.chem.qmw.ac.uk/iubmb/) Nomenclature of Carbohydrates (McNaught, 1997) specifies an extended form, a condensed form and a short form for the representation of glycans. For example, the core pentasaccharide of *N*-glycan is represented in

Extended form:



$$\begin{array}{rl} \alpha\text{-D-Man}p\text{-}1 & \\ \downarrow & \\ 6 & \\ \beta\text{-D-Man}p\text{-}(1\rightarrow4)\text{-}\beta\text{-D-Glc}p\text{NAc-}(1\rightarrow4)\text{-}\beta\text{-D-Glc}p\text{NAc-}(1\rightarrow & \\ 3 & \\ \uparrow & \\ \alpha\text{-D-Man}p\text{-}1 & \end{array}$$

or:



$$\begin{array}{l} \text{a-D-Man}p\text{-}(1\text{–}6)+ \\ \qquad\qquad\quad | \\ \qquad\qquad\quad \text{b-D-Man}p\text{-}(1\text{–}4)\text{-b-D-Glc}p\text{NAc-}(1\text{–}4)\text{-b-D-Glc}p\text{NAc} \\ \qquad\qquad\quad | \\ \text{a-D-Man}p\text{-}(1\text{–}3)+ \end{array}$$

Biomacromolecules, by C. Stan Tsai
Copyright © 2007 John Wiley & Sons, Inc.

Condensed form:

```
Man(α1-6)
         |
         Man(β1-4)GlcNAc(β1-4)GlcNAc(β1-
         |
Man(α1-3)
```

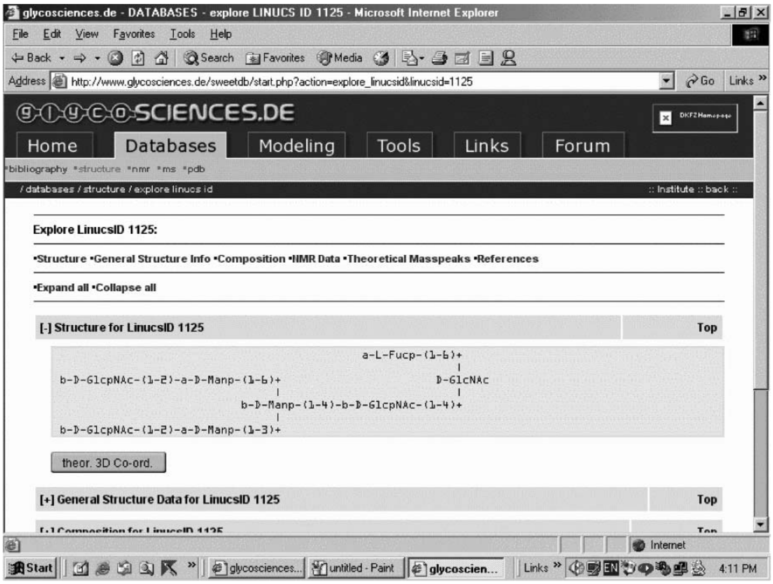
Short form: Manβ(Manα3)(Manα6)GlcNAcβ4GlcNAcβ-

These forms are commonly used in the representation of glycans, but none of them is particularly suited for computer processing because of the use of special symbols and some remaining ambiguities regarding the ordering of branches.

2. Linear notation for unique description of carbohydrate sequences (LINUCS) is developed by the German Cancer Research Centre (Bohne-Lang *et al.*, 2001) and is implemented for most of its Glycosciences Web (http://www.glycosciences.de/index.php) applications (Figure 17.1). The notation is derived in two steps:

1. application of syntax rule to perform a semantic transformation of carbohydrate graph into a framework resembles nested loops, which are then collapsed into a linear description containing nested brackets; and
2. application of the priority rules to create a unique notation.

The transformation can be slow and complicated with complex antennary glycans. A Web interface at http://www.glycosciences.de/tools/linucs/ is available for the conversion of the extended forms of IUPAC-IUBMB representations into LINUCS. For example, the LINUCS code for the core pentasaccharide is


glycosciences.de - DATABASES - explore LINUCS ID 1125 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.glycosciences.de/sweetdb/start.php?action=explore_linucsid&linucsid=1125
Go Links
GLYCOSCIENCES.DE
DKFZ Homepage
Home Databases Modeling Tools Links Forum
•bibliography •structure •nmr •ms •pdb
/ databases / structure / explore linucs id
:: Institute :: back ::

Explore LinucsID 1125:

•Structure •General Structure Info •Composition •NMR Data •Theoretical Masspeaks •References

•Expand all •Collapse all

[-] Structure for LinucsID 1125 Top

```
                                          a-L-Fucp-(1-6)+
                                                        |
b-D-GlcpNAc-(1-2)-a-D-Manp-(1-6)+                  D-GlcNAc
                                |                       |
                  b-D-Manp-(1-4)-b-D-GlcpNAc-(1-4)+
                                |
b-D-GlcpNAc-(1-2)-a-D-Manp-(1-3)+
```

theor. 3D Co-ord.

[+] General Structure Data for LinucsID 1125 Top

Internet
Start glycosciences... untitled - Paint glycoscien... Links 4:11 PM

Figure 17.1 Glycan search at SweetDB of Glycosciences
Glycan search *via* composition is initiated by entering Hex=3, HexNAc=4 and dHex=1. A list of candidate glycans is returned with options (buttons) for Explore, LiGraph, NMR and/or Pdb entries. Selecting the desired glycan by clicking Explore button opens the information page providing the structure, theoretical 3D coordinate, motif, structural data, NMR/MS peaks and/or PDB links.

[b-D-GlcpNAc]{[(4+1)] [b-D-GlcpNAc]{[4+1)][b-D-Manp]{[(3+1)]
[a-D-Manp]{}[(6+1)][a-D-Manp]{}}}}

3. LinearCode (Benin *et al*., 2002) represents a simpler linear representation of glycans that has been recently adapted by The Consortium for Functional Glycomics (http://www.functionalglycomics.org/). For example the core pentasaccharide is represented as

Ma3(Ma6)Mb4GNb4GMb1-

Instructions to the LinearCode can be accessed from http://www.glycomics.com/ and the use of the LinearCode to represent glycan sequences is described in subsection 6.1.3.

17.1.2 Glycobiology: Glycoforms

Glycobiology (Kobata, 1993; Dwek, 1996; Varki *et al*., 1999) deals with the structure and function of carbohydrates in biological systems. The biological information contained in the glycan structures is elucidated and used to understand biological behaviors and processes. Glycans appear as the most abundant and structurally diverse biomacromolecules that play a fundamental role in many important biological processes (Robyt, 1998; Dumitriu, 2005). They are commonly found in nature as glycoconjugates, which include glycolipids and glycoproteins. The chemical structure of glycoproteins allows recognition of two types of conjugates, the *O*-glycosidic and *N*-glycosidic glycoproteins (Corfield, 2000). The former carries the carbohydrate portion bound *O*-glycosidically to Ser or Thr, the latter *N*-glycosidically to the amide nitrogen in Asn with the N-X-S/T sequence, in which X can be any amino acid except Pro (Bause, 1983). Both groups differ not only in the type of linkage but also in the type of the bridgehead-saccharide. In the case of *O*-glycosidic glycoproteins, with rare exceptions. an α-D-GalNAc is always bound to the protein. The *N*-glycosidic glycoproteins all carry core pentasaccharide consisting of a *N*-acetylchitobiose (GlcNAcβ4GlcNAc) unit β-linked with the protein, on the 4-position of which is linked a β-Man*p* residue, which carries two α-Man*p* residues, (1→3)- and (1→6)-linked respectively. This inner core is found in all *N*-glycosidic glycoproteins.

Unlike proteins and nucleic acids, glycans are capable of forming many different combinatorial structures, including branched ones, from a relatively small numbers of sugar units. Each structure could potentially carry a specific biological message, thus widening the spectrum of reactivity that is possible from a limited number of monomers. The diversity of biological activities (bioactivities) is exhibited by a minimum change in number, linkage or branching of monosaccharide units in oligosaccharide chains of glycoproteins. A major function of protein-linked glycans is to provide recognition epitopes for protein receptors. For example, Table 17.1 illustrates sialic acid as the ligand in lectin recognition (Varki, 1997).

Most of oligosaccharides present in cells are attached to proteins or lipids. The majority of secreted and cell surface proteins of eukaryotes are glycosylated. These include proteins on the extracellular side of the plasma membrane, secreted proteins and proteins contained in body fluids. These glycoproteins are most easily accessible in the human body for diagnostic and therapeutic purposes. It is therefore not surprising that many clinical biomarkers and therapeutic targets are glycoproteins such as Her2/neu in breast cancer, prostate-specific antigen (PSA) in prostate cancer and CA125 in ovarian cancer. Glycosylation of proteins creates a set of glycoforms all of which share an identical polypeptide chain (identical amino acid sequence) but differ in oligosaccharide structures (saccharide sequences) or in disposition (selection, number and occupancy of glycosyla-

TABLE 17.1 Terminal oligosaccharide sequences recognized by some sialic acid binding lectins

Sialic acid containing Glycan sequence	Vertebrate			Insect	slug	Plant			Viral	
	Select	CD22	MAG	AA	LFA	MAA	SNA	TJA	InfA	InfC
NN[9T]a6AN-	–	–	–	–	–	–		–	–	+
NNa3Ab3(NNa6)ANb-	–	–	–	–	+	–	–	–	–	–
NNa3Ab3GNb-	–	–	+	–	+	–	–	–	+	–
NNa3Ab3(Fa4)GNb-	+	–		–	+	–	–	–		–
NNa3Ab4GNb-	–	–	+	–	+	+	–	–	+	–
NNa3Ab4(Fa3)GNb-R3	+	–		–	+	–	–	–		–
NNa6Ab4GNb-	–	+	–	+	+	–	+	+	+	–
NN[9T]a6Ab4GNb-	–	–	–		–	–			–	+

Notes: 1. Sialic acid has an uncompromising role in vertebrates as it is often at the capping position of cell surface glycan and is therefore exposed for interactions with exogenous cells and proteins. Sialic acid is a critical component of glycan epitopes responsible for initiating inflammatory leukocyte adhesion, B-cell signaling and activation and viral binding.

2. Terminal and minimum sialic acid (NN)-containing oligosaccharide sequences are shown in LinearCode. The continuing chains are either *O*-GalNAc-linked, or N-GlcNAC-/*O*-GalNAC-/ceramide-linked.

3. Abbreviations used: AA (Allo A-H) *Allomytina dichotoma* lectin; CD, cluster of differentiation antigen; infC, infA, influenza C hemagglutinin esterase; influenza A hemagglutinin; LFA, *Limax flavus* agglutinin; MAA, *Maackia amurensis* agglutinin; MAG, myelin-associated glycoprotein; select, selectins (C-type lectins); SNA, *Sambucus nigra* agglutinin; TJA, *Tricosanthes japonicum* agglutinin.

tion sites). Glycosylations are protein-specific, site-specific and tissue/cell-specific. Therefore a glycoprotein must be viewed as a collection of glycoforms. The observed changes in glycoforms are a reflection of the activity of the glycosyl transferases, which synthesize the oligosaccharides attached to the protein, and the glycosidases, which trim and degrade the glycan chains.

All glycoforms show the same amino acid sequence but differ in the location, number or sequence of attached glycan, and importantly their bioactivities. For example, bovine pancreatic ribonuclease (RNase) occurs naturally as a mixture of the unglycosylated protein (RNase A) and five glycoforms (RNase B, Man 5–9), in which 5–9 mannose residues are attached to the di-*N*-acetylchitobiose core. These glycoforms differ in their enzyme activities, i.e. decreasing activities in the order of RNase A > RNase ManO = RNase Man1 > RNase Man5 = RNase B (Rudd *et al.*, 1994).

The biosynthesis is ensured by highly specific transferases and is thereby genetically determined. A fully-built glycoconjugate always possesses the same conserved carbohydrate structures. Certain microheterogeneity can be attributed to a definite content of incomplete molecules. There are only a limited number (about 10) regularly occurring monosaccharides in glycoconjugates of higher animals. The apparently unlimited, documented structural variation arises solely through different branching possibilities, change of anomeric configuration, and also naturally through the variation of the sequences. Certain carbohydrate structures in glycoconjugates are definite unambiguous recognition features. The idea that oligosaccharides possess the possibility of conserving and transmitting information is stimulating and inspires the research in the newly created area of glycobiology. We endeavor to seek a code behind certain structural elements, comparable to the DNA code or the biological relevant tertiary structure in proteins determined by their primary structure. Carbohydrates possess the property of being hydrophilic. Thus in glycoproteins, a protein molecule is rendered more compatible with an aqueous medium through glycosylation. Oligosaccharides above all highly branched ones, have relatively rigid conformations, which are well adapted to place ionic groups and also lipophilic structural elements in certain positions on a protein surface or a membrane. In these cases the oligosaccharide serves as rigid framework for the installation of certain functional residues. Oligosaccharides can be voluminous and when densely distributed on the surface of a protein protect against denaturation as well as against attack of degrading enzymes.

Starting from the so-called bridgehead-structure, through further glycosylation, that is in part function- and organ-specific, the species-specific glycoproteins are formed. This apparently strict specificity is contradicted by the fact that identical structures arise from sources that exhibit no connection whatsoever and that a certain glycoprotein can also show, beside complete oligosaccharide, incomplete oligosaccharide structures. Though there are known examples in which both types occur (e.g. human plasminogen), usually a glycoprotein is either of the *O*- or *N*- glycosidic type. The appearance of so-called microheterogeneity is the incomplete biosynthesis of the finished glycoconjugate. The construction of *N*-glycosidic glycoproteins is not only the result of glycosylation but also of glycosidic cleavage, an occurrence that is designated as ‘processing’ or ‘editing’. This extraordinary complex processes accounts for all the different oligosaccharide structures found in glycoproteins. The complicated method of synthesis of the *N*-glycosidic oligosaccharide found only in eukaryotes supports the idea that there is a meaningful function for glycoconjugates. It seems improbable that all higher organisms produced the genetic, extraordinary lavish apparatus of the highly specific glycosyl transferases and the almost equally specific glycosidases, and developed the apparently uneconomic reaction pathways just to surround proteins and membranes with a hydrophilic coat. Probably the reason for the presence of these oligosaccharides is so that many different tasks may be accomplished. For one purpose, an precisely built oligosaccharide is required but for the others, any carbohydrate structure suffices. Future glycomics research may shed light to these important questions.

17.1.3 Glycomics: Response to post-genomic era

The term glycobiology describes the study of glycans and glycomics, in reference to genomic era and describes the study of the repertoire of glycans present among organisms and cell types. Automated methods for the isolation, amplification, sequence determination and expression profiling of nucleic acids have generated a vast amount of information. Genomics is concerned with the complete genomes of cells/organisms while proteomics describes an organism by the level of protein expression of active genomes. The development of proteomics has continued to exert dramatic impact on the study of diverse biological systems. One of the distinguishing features between genome and proteome in eukaryotic cells is that most of the gene products are posttranslationally modified and that glycosylation is a predominant modification. Whilst much of the concepts of genomics can be transferred to proteomics, it is difficult to apply similar principles to glycomics, since glycans are secondary gene products and their structure and function cannot be easily predicted from the DNA sequence.

The glycosylations that occur in a protein both during and after its expression has been implicated in protein function, and hence analyses of these modified forms of proteins, (i.e. glycoforms of glycoproteins) are essential to the understanding of the different protein expression patterns. For example, the A, B and O alleles of the genes for blood groups illustrate how the single nucleotide difference known as single nucleotide polymorphism (SNP) affects the glycan structure of glycoproteins. A and B alleles differ by four SNP substitutions. They code for related proteins that add different saccharide units to an antigen on the surface of red blood cells:

Allele	Sequence	Saccharide
A	· · · ·GCTGGTGACCCCTT· · · ·	*N*-acetylgalactosamine
B	· · · ·GCTCGTGACCGCTA· · · ·	galactose
O	· · · ·CGTGGT–ACCCCTT· · · ·	–

The O allele has undergone a mutation causing a phase shift, and produces no active enzyme. The red blood cells of type O individuals contain neither the A nor the B antigen. This is why people with type O blood are universal donors in blood transfusions. The loss of activity of the protein does not seem to carry any adverse consequences.

Thus one of the objectives of glycomics is to identify the glycoproteins that are markers of a particular phenotype, to characterize the specific changes in glycosylations that correlate with a change in phenotype and to target the metabolic step(s) that are responsible for this change. For example, agalactosylated glycoforms of aggregated IgG may induce association with the mannose-binding lectin and contributed to the pathology in rheumatoid arthritis (Rudd *et al*., 2001). The deglycosylated human chorionic gonalotropin (hCG) binds more strongly to target cells than natural glycosylated hCG but expresses no hormonal activity. Table 17.2 lists examples of altered glycoforms in human diseases.

Glycome describes not only the total cellular complement of ubiquitous, structurally diverse polymeric sugars and their conjugates (glycoproteins and glycolipids) but also the many genes and gene products involved in glycan synthesis/degradation, binding and regulation. Glycomics refers to the study of the complete carbohydrate complement of an organism (i.e. the sugar signature of various cells in an organism) that can be investigated by various approaches. The distinguishing features of glycomes are:

1. Glycans have a remarkably high information content owing to diversity in their primary chemical structures, which requires the application of various analytical techniques to discern the glycan fine structure and sequence. This information density derives from the fact that complex glycans contain not only information from diverse glycoses but also linkage variability.

TABLE 17.2 Examples of altered glycoforms associated with human disease

Disease	Glycoprotein	Alternation	Ref.
Cancer	Mucin of tumors	Over expression or exposure of specific short chain ***O***-linked glycans	1
Hepatic cancer	Serum α-fetoprotein	Different N-linked glycoforms	2
Inflammation, cancer, AIDS	Orosomucoid (α1-acid glycoprotein) in serum	Different N-linked glycoforms	3
Carbohydrate-deficient syndrome	Tranferrin, antithrombin, orosomucoid	Different N-linked glycoforms	4
Immune disorders	CD43 of T cells	Different O-linked glycoforms	5
Rheumatoid arthritis	IgG	Lowered terminal Gal in N-linked glycans	6
Schistosomiasis	Serum circulating antigens	Different O-linked glycoforms	7
Creutzfeldt-Jakob disease	Prion protein of cerebrospinal fluid	Different degree of glycosylation	8
Alcohol abuse	Serum transferrin	Desialylaion	9

Notes: 1. References quoted are: [1] Kim *et al*. (1997); [2] Taketa (1998); [3] Mackiewicz and Mackiewicz (1995); [4] Winchester *et al*. (1995); [5] Ellies *et al*. (1994); [6] Rudd *et al*. (2001); [7] Van Dam *et al*. (1996); [8] Furakawa *et al*. (1998); [9] Tagliaro *et al*. (1998).
2. Prion proteins cause a group of human and animal neurodegenerative diseases which manifest as infectious, genetic and sporatic disorder (Prusiner, 1996). Schistosomaiasis is a vascular parasitic disease caused by blood flukes of the genus *Schistosoma*.

2. Glycans, unlike DNA or proteins, are not synthesized *in vivo* by reading from a template. Complex glycans are created through the concerted action of several enzymes. This, together with a lack of proofreading machinery, results in heterogeneous and polydisperse glycoforms.

3. Glycans are synthesized in a template independent manner by the concerted action of glycosynthetic and glycolytic enzymes, thus there is no mechanism, at present, to amplify glycan structures that allows facilitated structure-function studies.

4. Glycans predominantly act at an extracellular and/or muticellular level. As such the screening of glycan activities using simple *in vitro* model systems does not always accurately correlate with the *in vivo* function. Furthermore, the glycan interactions typically involve the multivalent interactions of the glycan with the targets.

Glycan–protein interactions are viewed as important mechanisms for biological information transfer between cells and cell-substratum. Glycoproteins play crucial roles in biological processes as diverse as development, infection and immunity. One aspect of glycomic research concerns roles of glycans as recognition sites on cell surfaces, providing condensed information sources for various purposes. Thus there is a need to assemble a great deal of available knowledge about glycoproteins so that it can be used to help interpret the data acquired from glycomic analysis (Corfield, 2000), such as analysis of assemblage of glycan structures that change in normal and disease metabolism.

17.2 GLYCOMIC DATABASES AND SERVERS

17.2.1 Glycan structure

The advances in genomic and proteomic research have enabled many oligosaccharides of glycoproteins to be analyzed for their sequences, structures and functions. Thus the need exists for interfaces to bioinformatics tools constructed specifically for glycomes. The IUBMB site (http://chem.qmw.ac.uk/iubmb) offers general biochemical information for carbohydrates. General information, specifically keywords related to glycosciences including glycogene, glycoprotein, glycolipid, saccharide, glycotechnology and glycopathology, are available at GlycoWord (http://gak.co.jp/FCCA/glucoword/wordE.html). On-line glycan structure databases are listed in Table 17.3.

TABLE 17.3 Glycan structure databases

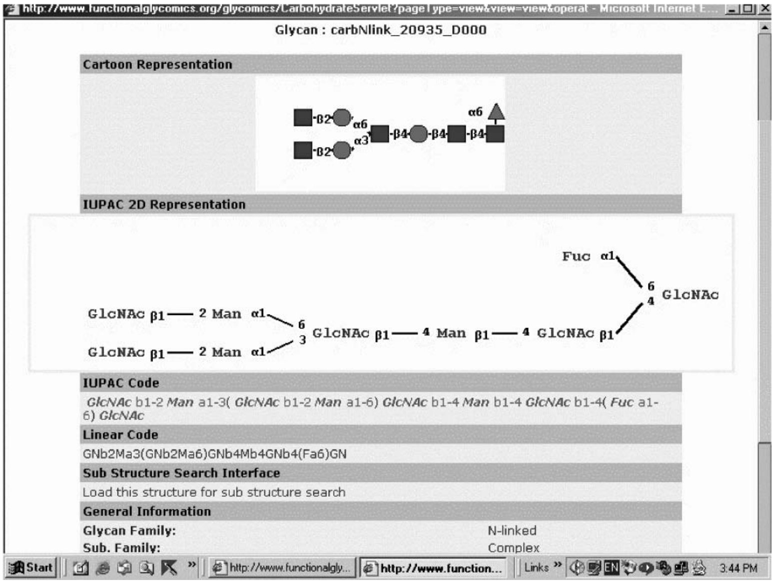
Database	URL	Description
IUBMB	http://www.chem.qmw.ac.uk/iubmb/	General structure information
CCSD	http://www.boc.chem.uu.nl/sugabase/carbbank.html	Prim structures, analysis
Glycosciences	http://www.glycosciences.de/index.php	Glycomic DB, modeling and tools
CSS	http://dkfz.de/spec/css/	Carbohydrate structures from PDB
Glycan DB	http://www.functionalglycomics.org/glycomics/molecule/jsp/carbohydrate/carbMoleculeHome.jsp	Glycan structures, search for substructures
GlycoBase	http://uslt.univ-lille1.fr/glycobase/	Structures of amphibian mucins
O-Glycobase	http://www.cbs.dtu.dk/databases/OGLYCBASE/	O-Linked GP oligosaccharides
SugaBase	http://www.boc.chem.uu.nl/sugabase/database.html	Prim structures, analysis
SweetDB	http://www.glycosciences.de/sweetdb/index.php	Annotation, Data collection

Notes: 1. CCSD and SugaBase have not been updated. However SugaBase which incorporates CCSD is still available.
2. Abbreviations used: GP, glycoproteins; prim, primary.

The Complex Carbohydrate Structure Database (CCSD), also known as CarbBank was originally created to provide literature, structure, analytical and biological information for complex carbohydrates including polysaccharides, glycolipids and glycoproteins (Doubet *et al*., 1989). The database is currently available at SugarBase and becomes part of SweetDB at Glycosciences (http://www.glycosciences.de/sweetdb/index.php). Glycosciences (http://www.glycosciences.de/index.php) maintained by the German Cancer Research Center (DKFZ) in Heidelberg, offers a wide range of glycomic database and tools:

1. *Databases* (http://www.glucosciences.de/sweetdb/index.php): The glycan structure and information can be searched and retrieved from the Structure database of SweetDB using substructure, exact structure, composition, molecular formula, classification and motif searches. Each candidate glycan is provided with options (buttons) for Explore (comprehensive exploration of structural information), LiGraph (display of structures), NMR (NMR spectrum/peaks), and/or pdb entries (list, links and explore glycan containing PDB). After selecting the desired glycan, the Explore option (button) offers information on molecular structure (IUPAC), structural data (formula, molecular weight, number/type of atoms), composition, display of 3D structure/download of the atomic coordinate (pdb format), NMR/ MS spectral peaks, link to the glycan containing PDB files and references. Independently, NMR and MS spectral peaks/profiles of a query glycan can be retrieved from the associated NMR and MS databases respectively. The Pdb database offers access to search/retrieve glycan components of glycoproteins archived in the PDB.
2. *Modeling* (http://www.glucosciences.de/modeling/index.php): The modeling service of SWEET2 (http://www.glucosciences.de/modeling/sweet2/doc/index.php) converts the glycan primary sequence into a 3D molecular template using the MM3 force field. Dynamic Molecules and Distance Mapping provide respective molecular modeling services of oligosaccharides. GlycoMaps DB (http://www.glucosciences.de/glycomapsdb/) is a database of glycan conformational maps and GlyProt simulates glycosylation of 3D proteins.
3. *Tools* (http://www.glucosciences.de/tools/index.php): Glycosciences offers a number of glycomic tools. These include Glycofragment (calculation and display of MS fragments), GlySeq (analysis the sequences around glycosylation sites), GlyTorsion (analysis of saccharide torsion angles derived from PDB), GlyVicinity (analysis of amino acids at the vicinity of saccharide residues), LiGraph (drawing of glycan structures), LINUCS (linear notation of glycans), pdb2linucs (extraction of saccharides from pdb files into LINUCS) and pdb-care (checking for carbohydrate structures in pdb files).

The Consortium for Functional Glycomics (http://www.functionalglycomics.org/) hosts three databases, Glycan Database, Glycan Binding Proteins (GBP database) and Glycosylation Pathways (GT database). The Glycan Database provides structural and chemical information as well as references for glycans. An interface search for sub-structure, molecular weight, and composition is available (Figure 17.2). The Protein Data Bank (PDB) at http://www.rcsb.org/pdb/ is also the primary database that provides 3D structure information for glycoproteins and some glycans. The 3D structure and sequence information of glycoproteins/carbohydrate-protein complexes (with cross-checked carbohydrate 3D structures) are extracted from the PDB by Carbohydrate Structure Suite (CSS) at http://www.dkfz.de/spec/css/ (Lütteke *et al*., 2005) and Pdb database of Glycosciences (http://www.glucosciences.de/sweetdb/start.php) or Pdb entries of SweetDB. A number of tools are also available at Glycosciences (http://www.glucosciences.de/tools/index.php) for extracting glycan structural information from the PDB. *O*-Glycobase


http://www.functionalglycomics.org/glycomics/CarbohydrateServlet?pageType=view&view=view&operat - Microsoft Internet E...
β2
β2
α6
α3
β4
β4
β4
α6
Fuc α1
6
4
GlcNAc
GlcNAc β1 — 2 Man α1
GlcNAc β1 — 2 Man α1
6
3
GlcNAc β1 — 4 Man β1 — 4 GlcNAc β1
Start
http://www.functionalgly...
http://www.function...
Links
EN
3:44 PM

Glycan : carbNlink_20935_D000

Cartoon Representation

IUPAC 2D Representation

IUPAC Code

GlcNAc b1-2 *Man* a1-3(*GlcNAc* b1-2 *Man* a1-6) *GlcNAc* b1-4 *Man* b1-4 *GlcNAc* b1-4(*Fuc* a1-6) *GlcNAc*

Linear Code

GNb2Ma3(GNb2Ma6)GNb4Mb4GNb4(Fa6)GN

Sub Structure Search Interface

Load this structure for sub structure search

General Information

Glycan Family:	N-linked
Sub. Family:	Complex

Figure 17.2 Retrieval of glycan search at CFG
A search for glycan is conducted at Glycan Database of Consortium for Functional Glycomics (http://www.functionalglycomics.org/glycomics/molecule/jsp/carbohydrate/carbMoleculeHome.jsp) using substructure, molecular weight, composition or linear nomenclature. The query return tabulates candidate glycans listing their molecular weight, IUPAC representation, composition, family (e.g. N-linked or O-linked), subfamily (e.g. core#, complex or hybrid) and sources. A selection of the desired glycan returns information page as shown.

(http://www.cbs.dtu.dk/databases/OGLYCBASE/), which uses neural networks to predict *O*-glycosylation sites, is the web database containing O-linked glycoprotein oligosaccharide information (Gupta *et al.*, 1999). Two glycomics tools are available at ExPaSy (http://www.expasy.org/tools), GlycanMass calculates oligosaccharide mass and GlycoMod predicts possible oligosaccharide structures on proteins.

17.2.2 Glycan analysis

There is a need for an integrated information package to be developed for glycoanalysis and similar databases or servers that provide biochemists with repositories of information and analysis tools (Table 17.4). Two web sites with structure information for conformational analysis of monosaccharides and disaccharides are maintained by Centre de Recherches sur les Macromolécules Végétales (CERMAV) at http://www.cermav.cnrs.fr. Molecular mechanics calculation of saccharide structure can be performed on-line at SWEET2 (http://www.glycosciences.de/modeling/sweet2/) (Bohne *et al.*, 1998) or off-line with Glydict (Frank *et al.*, 2002) provided by DKFZ. The SWEET2 site offers on-line service to convert the primary sequence of saccharide into a 3D structure template that can be optimized using the MM3 force field (Figure 17.3). The calculated 3D structure can be visualized on-line with An Interactive Server-side Molecular Image Generator (AISMAG) or downloaded as pdb file. An analysis of the conformational map is also available. Dynamic Molecules (http://www.mol-simulation.de/) provides an interactive molecular dynamic simulation and Distance Mapping (http://www.glycosciences.de/distmap/) offers computational methods for exploring the conformational space of oligosaccharides.

TABLE 17.4 Online servers for glycan analyses

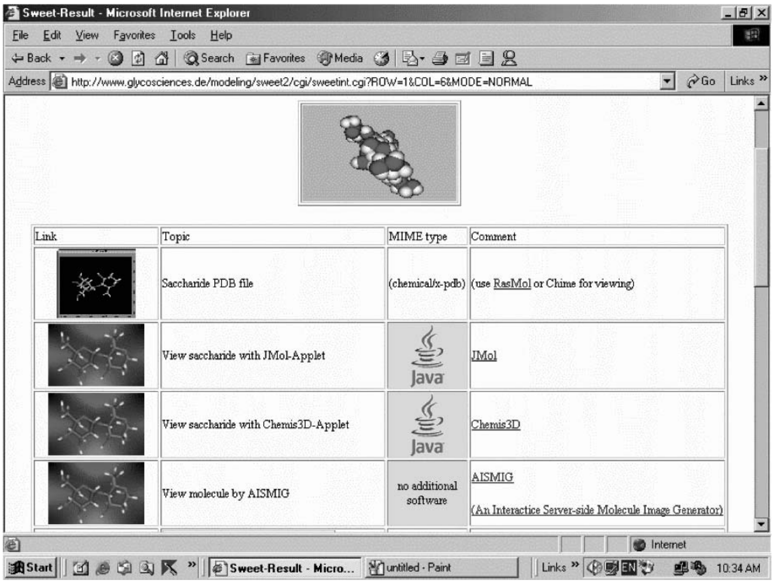
Server	URL	Description
Conformational analysis		
3D MS DB	http://www.cermav.cnrs.fr/cgi-bin/monos/monos.cgi	Conformations of monoS
3D DS DB	http://www.cermav.cnrs.fr/cgi-bin/di/di.cgi	Conformations of diS
Carp	http://www.glycosciences.de/tools/carp/	Generation of psi-phi plot
GlycoMaps	http://www.glycosciences.de/modeling/glycomapsdb/	Conformational maps of glycans
Dist. Mapping	http://www.glycosciences.de/distmap/	Conformational space
Dynamic Mol	http://www.mol-simulation.de/	Dynamic simulation
SWEET2	http://www.glycosciences.de/modeling/sweet2/	MM calculation, interface
Structure alignment, data mining		
GlySeq	http://www.glycosciences.de/tools/glyseq/	Sequences around gly sites
GlyTorsion	http://www.glycosciences.de/tools/glytorsion/	Glycan torsion angles from PDB
GlyVicinity	http://www.glycosciences.de/tools/glyvicinity/	Amino acids at the glycan vicinity
KCaM	http://glycan.genome.ad.jp	Structure search, alignment
Pdb2linucs	http://www.glycosciences.de/tools/pdb2linucs/	Glycan info from PDB
Pdb-care	http://www.glycosciences.de/tools/pdbcare/	Checking glycan residue in PDB
Spectral database		
CASPER	http://www.casper.organ.su.se/casper/	Simulation of NMR data
CCRC	http://www.ccrc.uga.edu/web/specdb/	NMR of xyloglycans
GlycoFragment	http://www.glycosciences.de/GlycoFragments/fragment.php4	Calculation of MS fragments
GlycoMod	http://expasy.org/tools/glycomod/	Interpretation of MS
Glypeps	http://dkfz-heidelberg.de/spec/glypeps/	Interpretation of MS
MS DB	http://www.glycosciences.de/sweetdb/ms/	Comparison, matching MS
NMR DB	http://www.glucosciences.de/sweetdb/nmr/	NMR spectra of oligoS
SPECARB	http://www.models.kvl.dk/users/	Raman spectra

Note: Abbreviations used are diS, disaccharides; gly, glycosylation; info, information; oligoS, oligosaccharides; MM, molecular mechanics; MS, mass spectra; monoS, monosaccharides and PDB, pdb file(s).

Conformational maps obtained from molecular dynamic simulation of glycans using the MM3 force field are available at GlycoMaps.

KEGG Carbohydrate Matcher (KCaM) at http://glycan.genome.ad.jp is a tool for the search, alignment and comparison of glycan structures via graphical interface. The search results are returned as a list of glycan structures in order of similarity based on an option for the local versus global alignment. The actual alignment can be viewed graphically along with annotated information (Aoki *et al*., 2004). Pdb2linucs of Glycosciences (http://www.glycosciences.de/tools/pdb2linucs/) offers a facility to detect carbohydrate compounds covalently attached glycans as well as noncovalently bound ligands, in PDB resource or a pdb file uploaded by the user. The carbohydrate compounds are identified and can be visualized using Chime plugin. For N- or O-linked glycans, the amino acid sequence of the respective protein is displayed with the glycosylation site highlighted (Lütteke *et al*., 2004).

Spectral databases of all known carbohydrate structures are undoubtedly useful for the identification of the carbohydrates at hand. Most of the oligosaccharide structures contained in the SweetDB are appended with ^{1}H- and/or ^{13}C-NMR spectra. Computer Aided Spectrum Evaluation of Regular Polysaccharides (CASPER) at http://www.casper.organ.su.se/casper/ is a tool for the analysis of the primary structures of oligosaccharides and for polysaccharides with repeating units based on NMR spectroscopy


Sweet-Result - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.glycosciences.de/modeling/sweet2/cgi/sweetint.cgi?ROW=1&COL=6&MODE=NORMAL
Go Links
Internet
Start Sweet-Result - Micro... untitled - Paint Links 10:34 AM

Link	Topic	MIME type	Comment
	Saccharide PDB file	(chemical/x-pdb)	(use RasMol or Chime for viewing)
	View saccharide with JMol-Applet	Java	JMol
	View saccharide with Chemis3D-Applet	Java	Chemis3D
	View molecule by AISMIG	no additional software	AISMIG (An Interactice Server-side Molecule Image Generator)

Figure 17.3 Molecular modeling of glycan with SWEET2
Two input formats, beginner and expert, are available on the SWEET2 home page (http://www.glycosciences.de/modeling/sweet2/). For a beginner, enter (select) a string of monosaccharides and linkages (e.g. b-D-Glc*p*2NAc(1-6)-a-D-Man*p*(1-4)-a-D-Man*p*(1-6)-b-D-Glc*p*2NAc) and Send. The return includes the 3D structure in different MIME types (file formats/molecular graphics) and options for optimization and conformational map (iterative) as shown.

(Stenutz *et al*., 1998). Based on information on components and linkages of glycans, NMR spectra are simulated for all possible structures of an oligosaccharide or repeating unit in a polysaccharide. The experimental spectrum is subsequently compared by ranking each of the simulated spectra according to their fit to spectral data. GlycoFragments (http://www.glycosciences.de/spec/GlycoFragments/fragment.php4) calculates all theoretically possible fragments of complex glycans to support the interpretation of mass spectra. The MS database of Glycosciences (http://www.glycosciences.de/sweetdb/ms/) compares each peak of a measured MS with the calculated fragments of all structures contained in the SweetDB (Lohmann and von der Lieth, 2004). GlycoMod and GlyPeps are web servers for the determination of glycosylation sites from MS by comparing experimentally measured masses to calculated peptide masses.

17.2.3 Glycosylation of proteins

Various related tools/approaches are employed to investigate glycomes. Some of web resources in support of such endeavors are listed in Table 17.5. GlyProt (http://dkfz-heidelberg.de/spec/glyprot/php/main.php) performs an *in silico* glycosylation of proteins with input 3D structures. The other two databases maintained by the Consortium for Functional Glycomics are Glycan Binding Proteins (GBP) and Glycosylation Pathways. The former (GBP) provides integrated presentation of the glycan binding proteins via interfaced search for all the publicly accessible data. The latter (GT) offers a graphical interface for navigation of the glycoenzyme database. The Technical University of Denmark hosts NetOGlyc, NetNGlyc and YinOYang for online prediction of mucine type *O*-glycosylation sites,

TABLE 17.5 Online utilities for glycoanalysis

Utility	URL	Description
Genes and proteins		
Animal lectin genes	http://ctld.glycob.ox.ac.uk/	Animal lectin genomics resources
BPGD	http://www.microbio.usyd.edu.au/BPGD/	Bacterial PolyS genes
3D Lectin DB	http://www.cermav.cnrs.fr/lectines/	3D structures of lectins
CAZY	http://afmb.cnr-mrs.fr/CAZY/	Carbohydrate active enzymes
GBP	http://www.functionalglycomics.org/glycomics/moleculle/jsp/gbpMolecule-home.jsp	Glycan binding proteins
Prediction of glycosylation sites		
Glycosyl pathways	http://www.functionalglycomics.org/static/gt/gtdb.shml	Navigation of glycoenzymes
GlyProt	http://www.dkfz-heidelberg.de/spec/glyprot/php/main.php	*In silico* glycosylation of proteins (input 3D)
NetNGlyc	http://cbs.dtu.dk/services/NetNGlyc/	O-linked GS
NetOGlyc	http://cbs.dtu.dk/services/NetOGlyc/	N-linked GS
YinOYang	http://cbs.dtu.dk/services/YinOYang/	β-GlcNAc-linked GS

Note: Abbreviations used: GS, glycosylation sites; polyS, polysaccharides.

N-glycosylation sites in human proteins and *O*-β-GlcNAc/phosphorylation sites respectively. CAZY (http://afmb.cnr-mrs.fr/CAZY/) is a comprehensive database for carbohydrate active enzymes (CAZYmes). CAZYmes are classified into sequence-derived families (Davis and Henrissat, 2002). They are modular, consisting of one or more catalytic domains in harness with many noncatalytic modules, which often posses a carbohydrate binding functionality. Active-site residues, molecular mechanisms and 3D structures are all conserved within families.

17.3 GLYCOMICS: GENETIC APPROACHES

Glycan production and modification comprise an estimated 1% of genes in the mammalian genome. Many of these genes encode for glycosyl-transferases and glycosidases, which are involved in construction of the glycan repertoire. The genes involved in the addition of the sugar portions of glycoconjugates are generically named glycogenes, which include the genes for:

- glycotransferases;
- glycolytic enzymes;
- glycose nucleotide synthetases;
- glycose nucleotide transporters, and in a broader sense, saccharide chain recognizing molecules such as:
    - lectins; and
    - saccharide chain receptors.

A single glycan of glycoconjugate is the joint product of dozen of genes. A comprehensive search for glycogenes in the genome and cDNA database to construct glycogene library would advance the knowledge of glycan structures and functions (Narimatsu, 2004).

Glycosynthetic and glycolytic genes families with overlapping activities but unique expression patterns are common, and glycan structures can be rapidly altered by these enzymatic activities. The regulated expression of glycosynthetic and glycolytic genes among cell types is primarily responsible for the dynamic complexity of glycan structures. The transfection of these glycogenes into cells leads to modification of the structure and function of the glycoproteins resulting in changes in glycome patterns. This enables the genetic manipulation and investigation of glycan structure and function (Lowe and Marth, 2003). Increasingly, examples of genetic alternations in glycan structure and expression have provided useful information about glycan function and etiology of glycan related diseases, such as human genetic defects in glycan formation (Table 17.6).

TABLE 17.6 Examples of human genetic defects in glycan formation

Enzyme affected by genetic defect	Phenotype associated with enzyme deficiency	Remark
Blood group GTs	Blood group polymorphism	Polymorphism in ABO bgl (Aa3T), H & Secrtor bgl (Fa2T), Lewis bgl (Fa3/4FT), P bgl (AT)
DPO synthesizing enzymes	CDG-Is	Mutations in gene(s) required for *N*-glycan synthesis characterized by varying degrees of mental and psychomotor retardation, coagulopathies and gastrointestinal signs and symptoms. CDG-Is are defects in enzymes involved in the synthesis of the pentasaccharide core and CDG-IIs are defects in extending the core of *N*-glycans.
GNT, Gase, Ab4T, F-transporter	CDG-IIs	
Ub4/GNa4 hs copolymerase	Inherited multiple exostoses	
Putative GTs	CMD	
AK, UDP-AT, UDP-A epimerase	Galactosemia	Impairment of UDP-A synthesis or accumulation characterized by normal infancy, hepatomegaly, and cataracts.
Ab4T-7	Ehlers-Danlos syndrome	Defect in glycosaminoglycan elongation characterized by connective tissue abnormalities.
Core 1 *O*-glycan Ab3T	Tn syndrome	Absence of Core I *O*-glycan characterized by thrombocytopenia, leukopenia and au autoimmume hemolytic anemia caused by auto-anti-Tn antibodies.
GNT-II	HEMPAS	Substantially reduced activity of GNT-II responsible for ineffective erythropoiesis, multinucleated erythroblasts, anemia
Agalactosyl glycoforms of IgG	Rheumatoid arthritis	Gal-deficient *N*-glycans in serum IgG causing complexation with ManBL that initiates/ perpetuates synovial inflammation.
O-Glycan abnormality on IgA	IgA nephropathy	
GDP-Man-4,6-dehydratase	Leukocyte adhesion deficiency	Recurrent infections and severe mental and growth retardation due to lack of GDO-Fuc formation

Notes: 1. Taken from Lowe and Marth (2003).

2. LinearCode is used for glycoses, e.g. Aa3 for $\alpha1\rightarrow3$Gal*p*, UbaT for $\beta1\rightarrow4$ Glc*p*A.

3. Abbreviations used: bgl, blood group locus (loci); CDG, congenital disorders of glycosylation; CMD, congenital muscular dystrophy; DPO, dolichol phosphate-oligosaccharide; Gase, glycosidase; GT, glycosyltransferase(s); HEMPAS, hereditary erythroblastic multinuclearity with positive acidified-serum lysis test; hs, heparan sulfate; T, transferase(s), e.g. Fa3/4T for $\alpha1\rightarrow3/1\rightarrow4$fucosylyltransferases.

A large number of glycosynthetic and glycolytic genes have been cloned. A strategy to analyze the functional properties of a gene is to knock out the gene or to overexpress it. While functions of some of these glycogenes have been rationalized by their pathophysiological effects associated with the gene knock-out or expression deficiencies (Table 17.1), the glycogene overexpression now becoming possible offers an efficient alternative to the knock-out or deletion approach for the identification of target molecules. For example, *N*-acetylglucosaminyltransferase III (GnT III) expression has been shown to correlate with an increased synthetic level for the bisecting GlcNAc and suppressed formation of β1→6 tri- and tetra-antennary *N*-glycans (Taniguchi, 2001). The experiment also identifies γ-glutamyltranspeptidase, E-cadherin, hyaluronate receptor CD44 and apolipoprotein B100 as endogenous acceptor molecules of GnT III, an involvement of GnT III in the down-regulation of the expression of the Gal epitope, epidermal growth factor receptor function and biosynthesis of pathogenic prions.

17.4 GLYCOMICS: PROTEOGLYCOMIC APPROACHES

17.4.1 Characterization of glycosylation sites

The glycosylation sites of N-Linked glycopeptides in complex biological samples are generally identified by two methods, both involving the immobilization of glycopeptides on a solid support and subsequent release of N-linked glycopeptides. MS is then used to identify the N-linked glycosylation sites and quantify the relative abundance of glycopeptides using isotope-coded tags. The two methods differ in the mechanism by which the glycopeptides are captured and tagged:

1. *Hydrazide covalent conjugation* (Zhang *et al*., 2003): Periodate oxidation converts the *cis*-diols of carbohydrates to aldehydes, which form covalent hydrazone bonds with the hydrazide (or amine) groups immobilized on the solid support. After washing/removing the nonglycosylated proteins, the immobilized glycoproteins are proteolyzed on the solid support. Nonglycosylated peptides are again washed away while the glycosylated peptides remain on the solid support. The α-amino groups of the immobilized glycopeptides are then labeled with isotopically light (H) and heavy (D)-succinic anhydride after lysine residues have been converted to homoarginine termed stable isotope tagging. The *N*-glycosylated peptides are finally released from the solid support using peptide-*N*-glycosidase F (PNGF). The released peptides are identified/quantified with MS–MS. Figure17.4 shows a schematic representation of hydrazide covalent isotope tagging.
2. *Lectin affinity capture* (Kaji *et al*., 2003): Glycoproteins are first purified from a biological protein mixture using lectin affinity chromatography. The recovered glycoproteins are digested with trypsin and the same lectin column captures glycopeptides. The N-linked glycopeptides are then cleaved with PNGF in $H_2{}^{18}O$. Since PNGF is an amidase, this enzymatic treatment also converts the glycosylated Asn to Asp, thereby incorporating ^{18}O form $H_2{}^{18}O$ into the glycosylation-site Asp to produce the specific tag termed isotope-coded glycosylation-site-specific tagging (IGOT). The ^{18}O-tagged peptides are analyzed by LC-tandem MS. The 2 Da shift introduced by PNGF-mediated reactions in $H_2{}^{16}O$ versus $H_2{}^{18}O$ can be used for quantitative profiling of glycoproteins. The released oligosaccharides can be collected for MS analysis. Figure 17.5 depicts schematically IGOT procedure via lectin affinity capture.

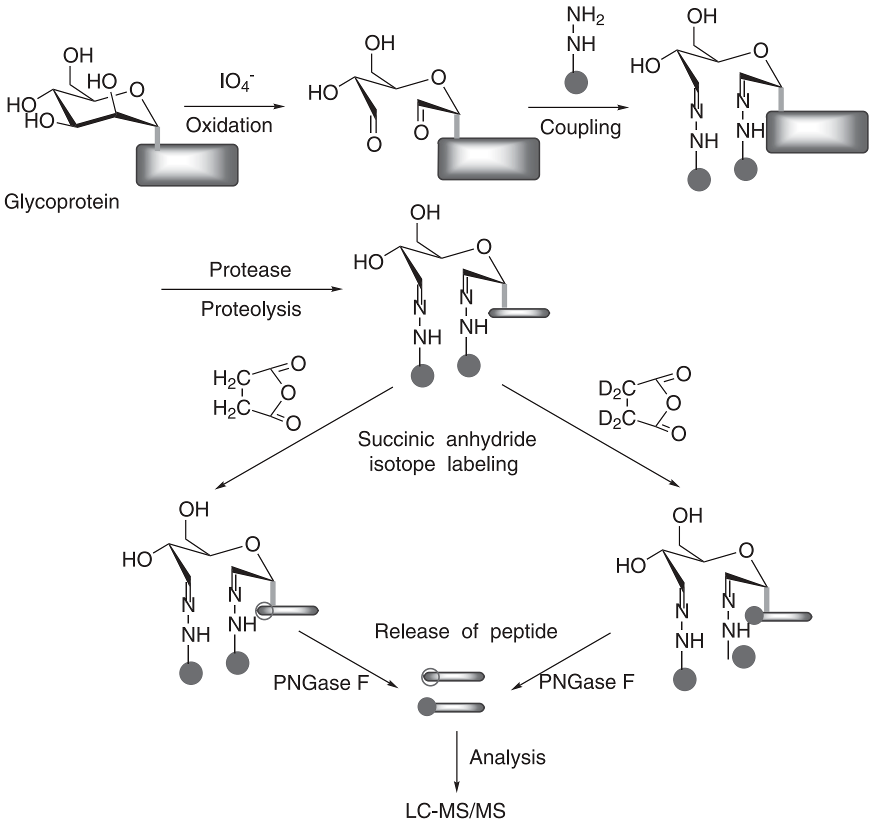

Glycoprotein
IO_4^-
Oxidation
NH_2
NH
Coupling
Protease
Proteolysis
Succinic anhydride isotope labeling
Release of peptide
PNGase F
PNGase F
Analysis
LC-MS/MS

Figure 17.4 Identification of N-linked glycosylation site by stable isotope tagging
The tagging process includes: (1) periodate oxidation of glycols to dialdheydes, (2) coupling of oxidized glycoprotein to hydrazide/amine solid support, and removing nonglycosylated proteins by washing, (3) proteolysis of the immobilized glycoproteins and removing nonglycosylated peptides by washing, (4) isotope labeling of α-amino groups of the glycosylated peptides with d_0- (H-) and d_4 (^{2}H-)-succinic anhydride, (5) release of the H- and ^{2}H (D)-glycosylated peptides as H- and D-peptides with PNGase F and (6) purification (microcapillary HPLC) and analysis of H- and D-peptides by tandem MS.

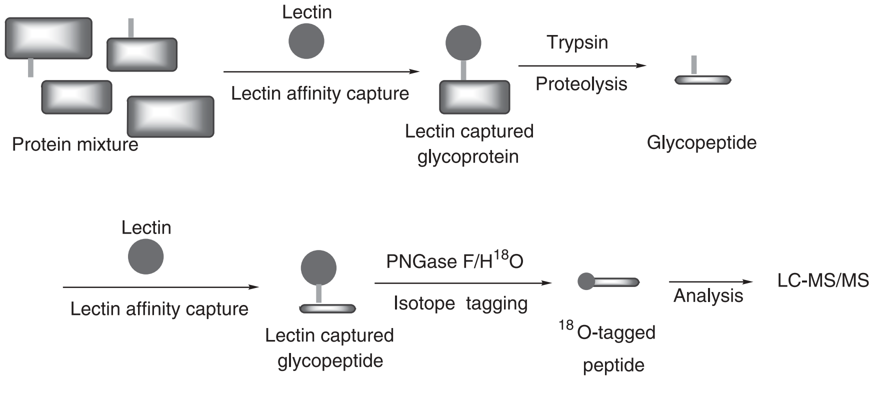

Lectin
Lectin affinity capture
Protein mixture
Lectin captured glycoprotein
Trypsin
Proteolysis
Glycopeptide
Lectin
Lectin affinity capture
Lectin captured glycopeptide
PNGase F/H^{18}O
Isotope tagging
^{18}O-tagged peptide
Analysis
LC-MS/MS

Figure 17.5 Identification of N-linked glycosylation site by IGOT
The IGOT process includes (1) purification of glycoproteins by lectin affinity chromatography, (2) proteolytic cleavage of lectin-captured glycoproteins to glycopeptides, (3) lectin capturing of glycopeptides, (4) release and isotope tagging (^{18}O labeling) of peptides, and (5) analysis of tagged peptides by tandem MS.

Mild β-elimination followed by replacement of the carbohydrate with dithiothreitol (DTT) through the Michael addition has been developed for the identification of protein-*O*-GlcNAc modification sites (Wells *et al*., 2002). The introduction of the DTT group can be used for affinity enrichment of the tagged peptides using an activated thiol-Sepharose column before their analysis by MS–MS. In addition, the two different isotopic compositions of DTT can be used to quantify *O*-GlcNAc-linked peptides. Because β-elimination followed by Michael addition has also been used to determine the sites of Ser/Thr phosphorylation, the determination of *O*-GlcNAc-linked peptides should be preceded by enzymatic removal of phosphate, immunoprecipitation with anti-*O*-pSer/pThr to remove *O*-phosphopeptides or with anti-*O*-GlcNAc antibody to purify *O*-GlcNAc-peptides.

17.4.2 Lectin and glycoenzyme-based proteoglycomics

Lectins that exist in most organisms are carbohydrate-recognition proteins mediating cell-discriminating phenomena. Their classification and structures have been described in subsection 10.5.1. Lectins are often complex, multidomain proteins, but carbohydrate-binding activity can usually be ascribed to a single protein module within the lectin polypeptide designated as a carbohydrate-recognition domain (CRD), each of which shares a pattern of invariant and highly conserved amino acid residues at a characteristic spacing.

Primary and 3D structures of a large number of lectins have been elucidated and are accessible at http://cermav.cnrs.fr/. In spite of the lack of primary sequence similarities, remarkable similarities have been noticed between the tertiary structures of lectins from diverse sources (Loris, 2002). Common tertiary structure is referred to as the lectin fold consisting of an elaborate jelly-roll, derived from antiparallel β-strands, arranged as two β-sheets in β-sandwich fold (e.g. galectins, pentraxins) or three β-subdomains with pseudo-threefold symmetry in β-trefoil fold (e.g. amaranthin, cysteine-rich domain of mannose receptor).

Mammalian lectins appear to be involved in recognizing structural diversity among the glycans and thereby modulate glycan-dependent molecular associations that can alter cellular physiology (Dodd and Drickamer, 2001). Because of their abilities to recognize/bind glycans specifically, lectins have been of value in establishing the presence or absence of specific glycan linkages and glycoforms in cells and organisms. For example, IGOT has been applied to the structural analysis of glycoproteins and glycan profiling (Hirabayashi, 2004).

The attachment of glycose residues is the most complicated co- and post-translational modification of proteins. Formation of the glycose-amino acid linkage is a crucial event in the biosynthesis of the glycan moieties of glycoproteins. In most instances, this step determines the nature of the glycose units that will subsequently be formed by cellular enzymatic machinery, which in turn influences the glycoprotein's biological activity. With the exception of the GlcNAcβ-Asn bond and the GPI anchor, the glycose-amino acid linkage is formed by the direct enzymatic transfer of an activated monosaccharide to a specific amino acid residue in the polypeptide chain (Table 17.7).

The core glycopeptide structure remains regardless of further steps in glycan linkage formation or subsequent hydrolytic trimming. *N*-Glycans are assembled in a step-wise manner by the sequential actions of glycosyltransferases and glycosidases. It is characterized by the assembly of a dolichol-linked oligosaccharide precursor that is transferred *en bloc* to a nascent polypeptide chain. The oligosaccharide chain is remolded by a series of reactions mediated by glycosyltransferases and glycosidases in a hierarchical order. *O*-Glycan chain initiation by the ppGalNAc transferases is followed by modification involving distinct branch-specific glycosyltransferases. Glycosidases are not known to be

TABLE 17.7 Linkages and enzymes involved in the synthesis of glycopeptide bonds

Linkage	Enzyme	Glycosyl donor	Consensus sequence or domain/module	Example
N-glycosyl				
GlcNAC-β-Asn	OligoT	DP-oligo	E-X′-S/T	Ovalbumin, fetuin, insR
Glc-β-Asn	GlcT	UDP-G	E-X′-S/T	Laminin, *H. halobium*
O-glycosyl				
GalNAc-α-Ser/Thr	GalNAcT	UDP-AN	Repeat domains with S,T,P,G,A	Mucin, glycophorin
GlcNAc-α-Thr	GlcNAcT	UDP-GN	T rich domain near P	
GlcNAc-β-Ser/Thr	GlcNAcT	UDP-GN	S/T rich domain near P,V,A,G	Nucl/cytopl proteins
Man-α-Ser/Thr	ManT	DP-M	S/T rich domain	Yeast mannoproteins
Fuc-α-Ser/Thr	FucT	GDP-F	EGF module (CXXGGT/SC)	CoagF and fibF
Glc-β-Ser	GlcT	UDP-G	EGF module (CXSXPC)	Coagulation factor
Xyl–β-Ser	XylT	UDP-X	S-G (A) near one/more acidic	Proteoglycan
Glc/GlcNAc-Thr	cytotoxin	UDP-G/GN	Rho: Thr-37[4]	Rho protein (GTPase)
Gal-Thr	GalT	UDP-A	G-Z-T	Vent worm collagen
Gal-β-Hyl	GalT	UDP-A	Collagen repeats, X-Hyl-G	Collagen, core lectin
Ara-α-Hyp			Repetitive Hyp rich, KPhPhPV	Potato lectin
GlcNAc-Hyp	GlcNAcT	UDP-GN	Skp1: Hyp-143[4]	*Dict* cytoplasmic protein
Glc-α-Tyr	Glycogenin	UDG-G	Glycogenin: Tyr-194[4]	Muscle/liver glycogenin
Phosphoglycosyl				
GlcNAc1P-Ser	GlcNAcPT	UDP-GN	S rich domain, ASSA	*Dict* proteinases
Man-α1P-Ser	MPT	GDP-M	S rich repeat domain	L. mexicana acid Pase
C-mannosyl				
Man-α-Trp	ManT	DP-M	W-X-X-W	RNase2, IL12, properdin

Notes: 1. Hyl (hK) and Hyp (hP) refer to hydroxylysine and hydroxyproline respectively. X = any amino acid; X′ = any amino acid except Pro; Z = A,P,R,S,Hyp.

2. Amino acids and glycoses: three-letter codes are used for linkages and enzymes in agreement with common representations while one-letter codes are used for glycosyl donors and consensus sequences.

3. Abbreviations used: coagF, coagulation factor; cytopl, cytoplasmic; *Dict*, *Dictyostelium*; DP, dolichol pyrophosphate; EGF, epidermal growth factor; fibF, fibrinolytic factor; IL, interleukin; insR, insulin receptor; nucl, nuclear; oligo, oligo-glycosyl or oligosaccharide; Pase, phosphatase; Rho, Rho family of small GTPases; Skp1 of Skp1-cullin-F-box; T, transferase.

4. The specific amino acid residue involved in the glycopeptide bond is given where glycopeptide bond appears to be limited to specific protein with established amino acid sequence.

required for *O*-glycan synthesis as they are in *N*-glycan synthesis. A hallmark of glycosyltransferases is their remarkable degree of substrate specificity. The substrate is typically the previous step in the pathway, and therefore absence of a single enzyme can restrict further diversification of the glycan repertoire as downstream steps cannot occur. This high degree of substrate specificity, in order to maintain a hierarchical order in assembling the glycan chain, is best illustrated, for example, by GlcNAc transferase isozymes (GlcNacT-I to -VIII), all of which add GlcNAc residue (from UDP-GlcNAc) but at different steps (different substrates) in *N*-glycan synthesis (Figure 17.6). Therefore genetic and biochemical (e.g. inhibition study) manipulations of glycosyltransferase activities should afford valuable information concerning glycan structures and functions (Park *et al*., 1996).

The use of specific exo- and endoglycosidases for the sequence determination of glycans has been described (subsection 6.2.3). Thus experiments can be designed to test the structural possibilities of complex glycans. Glycotransferases and glycosidases have been employed for enzymatic synthesis of oligosaccharides and can be used to remodel

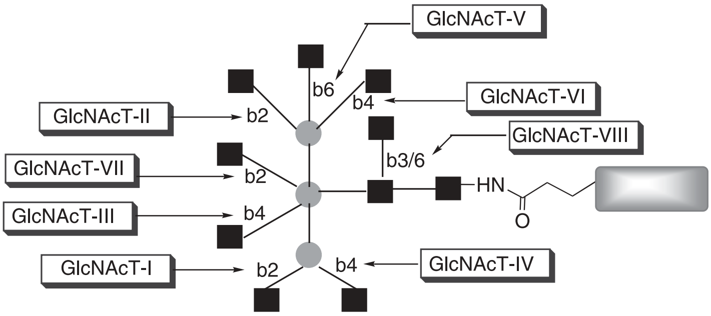


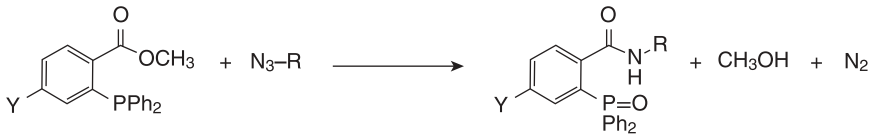
Figure 17.6 Action patterns of acetaminoglucosyl transferase isozymes
Acetaminoglucosyl transferases (GlcNAcTs) catalyze β-GlcNAc additions to form various glycosyl linkages (b for β-linkages) of N-glycan in a different hierarchical order as illustrated by the acetaminoglycosylation sites for the GlcNAc transferase (GlcNAcT) isozymes. GlcNAc and Man are represented by square and circle respectively.

glycan chains for the production of homogeneous glycans/glycoproteins (Hanson *et al*., 2004). Modification of the recombinant proteins with glycoenzymes is an attractive avenue to produce homogeneous glycoproteins for functional studies.

17.4.3 Metabolic oligosaccharide engineering

Cellular studies of glycan functions require techniques that allow manipulation of glycans within their native environment. Metabolic oligosaccharide engineering (Dube and Bertozzi, 2003; Keppler *et al*., 2001) permits the introduction of subtle modification into monosaccharide residues within cellular glycans enabling delineation of the molecular basis of their function. Technically unnatural monosaccharides are supplied as metabolic precursors to growing cells, and are transformed into activated nucleotide glycoses, incorporated via biosynthetic pathways into glycoconjuates, which are destined for secretion, delivery to cellular compartments or presentation on the cell surface (Figure 17.7A). The modification of the unnatural substrate is represented in its biosynthetic product, permitting investigation into the structure of a glycose with its cellular activity.

The incorporation of designed monosaccharides on the cell surface provides an avenue for the chemical remodeling of living cells. Unique chemical functional groups can be delivered to the cell surface glycan by bio-orthogonal chemoselective ligation (Figure 17.7B), for example between the *N*-α-azidoacetamidoglycoses of the cell surface and tagged phosphine compound via Staudinger ligation (Saxon and Bertozzi, 2000). The functionalized phosphine reacts with azide to form a stable amide bond:



Beyond its utility in elaborating cell surface glycan structures, chemical tagging of azide-labeled glycan via phosphine ligation permits the identification/purification of specific glycoproteins subtypes for proteomic and glycomic studies.

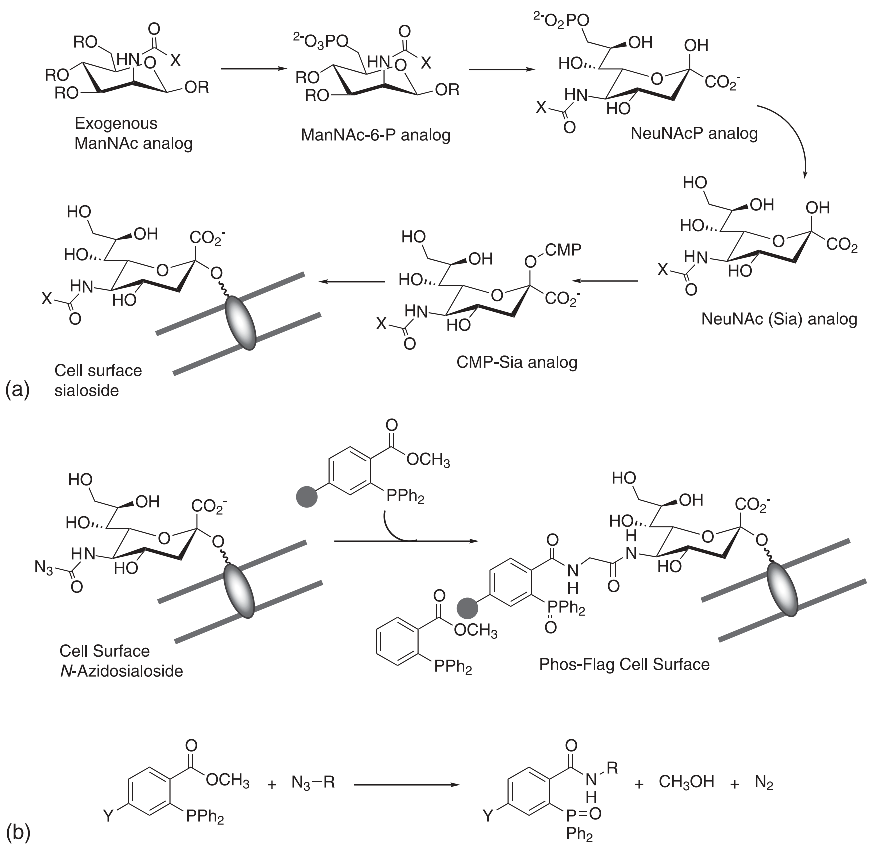


Figure 17.7 Metabolic oligosaccharide engineering
(A) Metabolic oligosaccharide engineering: Unnatural monosaccharides (e.g. N-x-acetylmannosamide where x = alkyl or azido groups) are supplied to growing cells/injected to animals, converted into appropriate glycose analogs (e.g. N-alkyl or azido-sialic acid) and incorporated into cell surface glycoproteins *via* series of tolerant biosynthetic reactions. (B) Chemical cell surface remodeling: Once displayed on the cell surface glycoproteins, for example the azides are poised for Staudinger ligation with phosphine compounds to generate labeled glycans (Phos-flag glycans). Adapted from Dube and Bertozzi (2003)

17.4.3 Recombinant glycoproteins

Expression systems vary in their ability to glycosylate proteins, depending on the cell line/culture conditions and consequently affect their functions. The absence or presence/extent of glycosylation modulates the effectiveness of some recombinant proteins. For examples, human lymphotoxin, thyroperoxidase and granulocyte colony-stimulating factors are recombinant proteins that require glycosylation for bioactivities. Other recombinant proteins that are glycosylated in their native forms, such as human growth hormone, interlukin-3 and thromopoietin are active even when nonglycosylated. Experimental approaches in which the *in vivo* changes of glycoproteins are monitored, can provide an understanding of the native glycoform heterogeneity and function.

The expression patterns of glycoproteins can be monitored by 2DE in conjunction with MS. A general method for locating the glycoproteins in a mixture of protein on a 2DE gel is a specific glycan staining. Sensitive lectin (Corfield, 2000) and hapten recognition protocols are available and several commercial products (e.g. Bio-Rad glycan detection kit or Boehringer-Mannheim DIG detection kit) can be employed. Capillary zone electrophoresis provides a useful approach in the high-throughput monitoring of recombinant glycoprotein isoforms (Susuki and Honda, 1998).

17.5 GLYCOMICS: CHEMOGLYCOMIC APPROACHES

17.5.1 Structural analysis of glycans

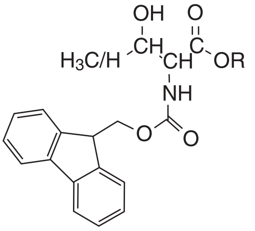
It is estimated that more than 50% of human proteins are glycosylated. However, our knowledge concerning the effect of glycosylation on the structure and function of proteins is limited because of difficulties in obtaining structural and therefore functional information for structurally diverse glycoproteins. The general strategy for the sequence determination of glycans has been described in subsection 6.2. There is no single analytic method that can yield glycan structures of glycoproteins. In order to obtain structural information of glycans, various glycoforms have to be separated/purified, commonly by the use of 2D gel electrophoresis (GE) or HPLC. Glycans are then released from glycoproteins and captured/purified.

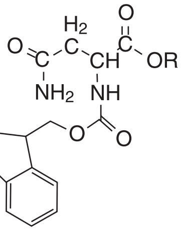
Mass spectrometry (MS) has become the method of choice for carrying out structure determination of glycans (Dell and Morris, 2001). Three technologies, fast atom bombardment (FAB), electrospray ionization (ESI) (Fenn *et al.*, 1989) and matrix-assisted laser desorption ionization (MALDI) (Kaufmann, 1995), permit the direct ionization and desorption of nonvolatile biomolecules and are applicable to glycan analyses (Kuster *et al.*, 1997; van den Steen, 1998). Glycan sequence can be deduced from knowledge of the structure-specific mass spectrometric fragmentation patterns (glycan mass fingerprinting). Tandem MS is able to determine the possible monosaccharide composition of the different glycoforms and provides information on the sequence and linkage of the residues by the characteristic way in which they fragment by collision-induce dissociation. Fragmentation of the glycosidic bonds occurs with post-source decay MALDI-TOF and in-source ESI-CID-TOF. However, complete characterization of the structure of the glycoforms requires more than just measurement of the mass of the sugars. Compared with amino acid masses used in peptide mass fingerprinting, there are many glycose residues that have the same molecular mass, and linkage and sequence cannot be determined easily by MS alone. Nevertheless, glycoproteins found within a particular tissue and/or species will restrict the possible glycan structures that can account for a specific mass. In general, the MS detection systems are coupled with enzyme digestion, NMR spectroscopy and/or chemical derivatization for complete glycan structure analysis and profiling.

17.5.2 Glycoprotein syntheses in glycomics

Glycoproteins are typically expressed as a mixture of glycoforms, their oligosaccharides being generated by a template-independent biosynthetic process. Whilst investigation of their function requires sufficient quantities of homogeneous glycans and glycoproteins, the isolation of pure glycans and glycoproteins from natural sources is extremely difficult owing to their structure complexity and isoforms. Thus an access to homogeneous glycans and glycoproteins relies on synthetic approaches (Davis, 2002; Grogan *et al.*, 2002; Hölemann and Seeberger, 2004). Chemical synthesis can aid in procuring structurally defined glycans and glycoproteins in sufficient amounts for functional and therapeutic studies.

General approaches to the chemical synthesis of glycans have been described in Chapter 8 and in particular the preparation of oligo- and polysaccharides by the solid phase and enzymatic syntheses in the subsection 8.4.3. Therefore we will focus our attention on the synthesis of glycoproteins here. Synthesis of glycopeptides and glycoproteins is commonly accomplished by incorporation of a suitably protected *O*-/*N*-glycosyl amino acid into a polypeptide by solid-phase synthesis (SPS). Fluorenyl-9-methoxycarbonyl (Fmoc) protected amino acids are preferred amino acid acceptors for SPS, since the removal of base-labile Fmoc is compatible with the presence of acid-sensitive glycosidic bonds.





The hydroxyl groups of the carbohydrates are protected as acetyl or benzoyl ester which, upon cleavage of the completed glycopeptide from the solid support, are readily removed by treatment with sodium methoxide or hydrazine.

Typically the complex *O*-glycan structures are constructed after formation of the desired α-glycosyl-Ser/Thr via glycosylation of protected Ser/Thr with orthogonally protected glycosyl halide/thioether referred to as a α-*O*-linked cassette (Figure 17.8). Orthogonally protected glycosyl halide/thioether is coupled with protected Ser/Thr. By orthogonally removing protecting groups on the C-3, C-4 and C-6 hydroxyl groups of the O-linked cassette, the controlled sequential addition of branching sugars can be performed to yield core O-linked glycans. A similar strategy can be adapted for the synthesis of complex N-linked glycans via β-glycosyl-Asn termed N-linked cassette.

The glycosyl-Fmoc-Ser/Thr or glycosyl-Fmoc-Asn is readily introduced into the elongating peptide chain via SPS. The development of an automated solid-phase oligosaccharide synthesizer (Plante *et al.*, 2001) should improve the accessibility of complex glycans of biological relevance. An elongation of the solid-phase synthesized peptide chain of glycopeptides can be accomplished by chemical ligation (subsection 8.4.1). The use of expressed peptides with an intein in the chemical ligation termed expressed protein ligation (EPL) has achieved the semisynthesis of full length proteins with complex but define glycans (Tolbert and Wong, 2000).

An approach to bypass cumbersome stereoselective glycosylation and protecting strategies for building complex glycan structures is by the use of hosts of glycosyltransferases and glycosidases in the enzymatic synthesis (Leppanen *et al.*, 2000). This can be accomplished by either: i) immobilized substrate with flow-through enzymes; or ii) immobilized enzymes with flow-through reactants. Various chemoenzymatic strategies have been applied to synthesize glycans and glycoproteins (Hanson *et al.*, 2004). Nonsense suppressive mutagenesis (subsection 8.6.2) using glycosylated aminoacyl-tRNA such as glycosylated Ser/Thr-tRNA may offer a methodology for the synthesis of O-linked glycoproteins.

17.5.3 Glycochip

The technology common to DNA and proteins has become realized in the form of biochips, which include carbohydrate arrays or glycochips. The emergence of biochips has revolu-

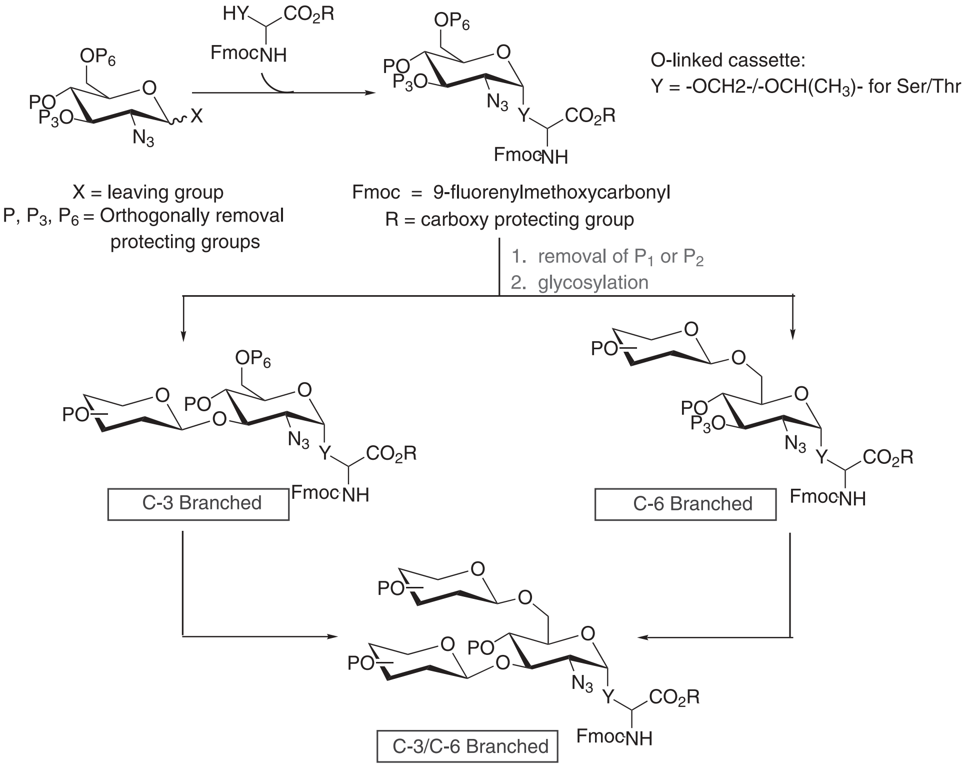
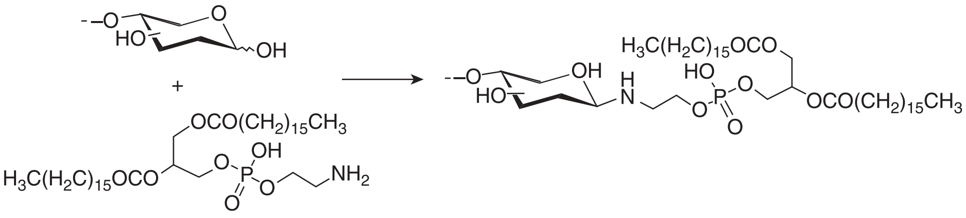

HY
CO2R
FmocNH
OP6
PO
P3O
O
X
N3
Y
O-linked cassette:
Y = -OCH2-/-OCH(CH3)- for Ser/Thr
X = leaving group
P, P3, P6 = Orthogonally removal protecting groups
Fmoc = 9-fluorenylmethoxycarbonyl
R = carboxy protecting group
1. removal of P1 or P2
2. glycosylation
C-3 Branched
C-6 Branched
C-3/C-6 Branched

Figure 17.8 Synthesis of complex O-linked glycosyl amino acids *via* Ser/Thr cassettes
For the α-O-linked cassette, the glycosylation of Fmoc-Ser/Thr-ester is carried out by the use of 2-azido halo- or thioglycoside donor (X = Br, Cl, SC_2H_5) to ensure high α-selectivity. Selective removal of the orthogonal protecting groups permits the sequential introduction/extension of glycan branching/chain. The 2-azido group is converted to 2-acetamido group *via* hydrogenation and N-acetylation. For the β-N-linked cassette, the glycosylation of Fmoc-Asn is carried out by the use of 2-acetamido aminoglycoside (2-acetamido-glycosylamine which can be produced by the condensation of a free reducing glycose with ammonium carbonate). The N-linked acetamidoglucosyl-Fmoc-Asn is converted to pentasaccharyl Fmoc-Asn. The resulting cassettes are used in SPS.

tionized the high-throughput screening (HTS) techniques and functional profiling of biomacromolecules. In biochip HTS, fluorescence-tagged probe molecules are incubated with biochip with immobilized targets. Unbound probe molecules are washed away and the binding ensemble is determined via fluorescence of the bound probe. For studying glycan-ligand interactions, the presentation of glycans on a surface is advantageous because glycan interaction in biological systems often occur on surfaces and are multivalent in nature (Kiessling *et al*., 2000).

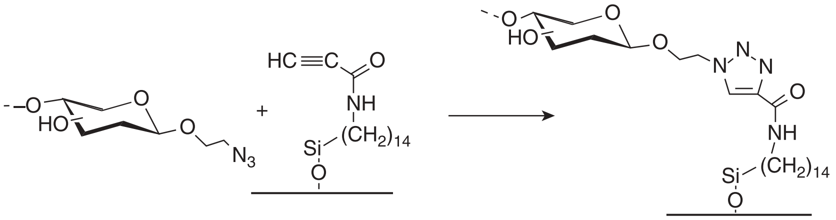
General procedures to manufacture, application and management of microarrays have been described in section 14.3. Special concerns to prepare carbohydrate microarrays will be dealt with in this subsection. Two methods are used to immobilize glycans on the surface for the fabrication of glycochips (Feizi *et al*., 2003): noncovalent and covalent immobilization technologies.

17.5.3.1 Non-covalent immobilization. Unmodified glycans of high molecular weight (M_r > 20 kDa) can be immobilized nonspecifically on nitrocellulose-coated glass slides. Since the immobilization involves the adsorption of glycans on an hydrophobic surface, a greater affinity is achieved with longer saccharide chains. Polysaccharides, glycosaminoglycans, glycoproteins and neoglycoconjugates have been immobilized to probe antibodies in the antigenic analyses. Oligosaccharides (containing 2–20 glycose units) and monosaccharides are linked to lipids via reductive amination to the amino phospholipid to yield neoglycolipids, which are immobilized by noncovalent absorption.

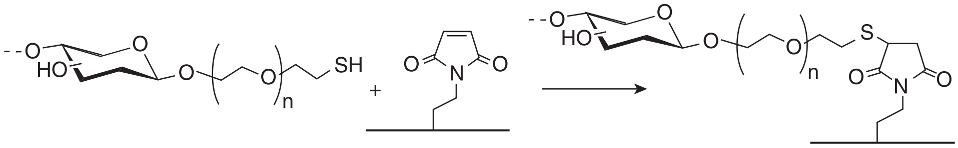


A 1,3-dipolar cycloaddition reaction between a carbohydrate bearing an azido group and an alkyne linker (optimal chain length of C_{13}–C_{15}) affords lipid-bearing glycans that can be immobilized on polystyrene microtiter plates. This approach (reverse sequence) can be adapted for the covalent immobilization technique. The main advantage of noncovalent immobilization is that the glycans do not need to be modified prior to immobilization. However, the buffers (limited to those that do not include detergents or nonaqueous solvents) and washing conditions must be carefully chosen so that the glycans remain bound to the surface of the glycochips.

17.5.3.2 Covalent immobilization. The covalent immobilization has been used to construct high-density DNA/RNA and protein microarrays. Chemical libraries constructed from combinatorial synthesis have been spatially arrayed on glass plates to facilitate screening of the library for members that bind tightly to proteins (MacBeath *et al*., 1999). In a reverse strategy of the noncovalent immobilization, glycans are attached to microarrays via a 1,3-dipolar cycloaddition reaction (Fazio *et al*., 2002) between the azido group of azidoalkyl glycoside and the alkyne group at the end of hydrocarbon arm of the microplate.



Maleimide-linked glycans to glass slides with thiol surface has been presented (Figure 14.4C). In an analogous approach, glycans with a polyethylene glycol (PEG) tail that terminated with thiol functionality on the reducing end are immobilized to the maleimide-activated solid surfaces.



Glycans immobilized covalently are stable under various conditions. These glycochips have been used for investigating glycan–protein interactions, HTS and lectin/glycoenzyme discovery.

17.6 REFERENCES

Aoki, K.F., Yamaguchi, A., Ueda, N. *et al.* (2004) *Nucleic Acids Research*, **32**, W267–72.

Bause, E. (1983) *Biochemistry Journal*, **209**, 331–6.

Benin, E., Neuberger, Y., Alshuler, Y. *et al.* (2002) *Trends in Glycoscience Glycotechnology*, **14**, 127–37.

Bohne, A., Lang, E. and von der Lieth, C-W. (1998) *Journal of Molecular Modeling*, **4**, 33–43.

Bohne-Lang, A., Lang, E., Föster, T., and von der Lieth, C-W. (2001) *Carbohydrate Research*, **336**, 1–11.

Corfield, A.P. (ed.) (2000) *Glycoprotein Methods and Protocols: The Mucins*, Humana Press, Totowa, NJ.

Davis, B.G. (2002) *Chemistry Review*, **102**, 579–601.

Davis, G.J. and Henrissat, B. (2002) *Biochemistry Society Transactions*, **30**, 291–7.

Dell, A. and Morris, H.R. (2001) *Science*, **291**, 2351–6.

Dodd, R.B. and Drickamer, K. (2001) *Glycobiology*, **11**, 71R–9R.

Doubet, S., Bock, K., Smith, S. *et al.* (1989) *Trends in Biochemistry Science*, **14**, 475–7.

Dube, D.H. and Bertozzi, C.R. (2003) *Current Opinions in Chemical Biology*, **7**. 616–25.

Dumitriu, S. Ed (2005) *Polysaccharide: Structure Diversity and Functional Versatility*, 2nd edn, Marcel Dekker, New York.

Dwek, R.A. (1996) *Chemistry Review*, **96**, 683–720.

Ellies, L.G., Jones, A.T., Williams, M.J. and Ziltener, H.J. (1994) *Glycobiology*, **6**, 885–93.

Fazio, F., Bryan, M.C., Blixt, O. *et al.* (2002) *Journal of the American Chemistry Society*, **124**, 14397–402.

Feizi, T., Fazio, F., Chai, W. and Wong, C-H. (2003) *Current Opinions in Structural Biology*, **13**, 637–45.

Fenn, J.B., Mann, M., Meng, C.K. *et al.* (1989) *Science*, **246**, 64–71.

Frank, M., Bohne-Lang, A., Wetter, T. and von der Lieth, C-W. (2002) *In Silico Biology*, **2**, 427–39.

Furakawa, H., Doh-ura, K., Kikuchi, H. *et al.* (1998) *Journal of Neurological Sciences*, **158**, 71–5.

Grogan, M.J., Pratt, M.R., Marcaurelle, L.A. and Bertozzi, C.R. (2002) *Annual Reviews in Biochemistry*, **71**, 593–634.

Gupta, R., Birch, H., Rapacki, K. *et al.* (1999) *Nucleic Acids Research*, **27**, 370–2.

Hanson, S., Best, M., Bryan, M.C. and Wong, C-H. (2004) *Trends in Biochemistry Science*, **29**, 656–63.

Hirabayashi, J. (2004) *Glycoconjugate Journal*, **21**, 35–40.

Hölemann, A. and Seeberger, P.H. (2004) *Current Opinions in Biotechnology*, **15**, 615–22.

Kaji, H., Saito, H., Yamauchi, Y. *et al* (2003) *Nature Biotechnology*, **21**, 662–7.

Kaufmann, R. (1995) *Journal of Biotechnology*, **41**, 155–75.

Keppler, O.T., Horstkorte, R., Pawlita, M. *et al.* (2001) *Glycobiology*, **11**, 11R–18R.

Kiessling, L.L., Gestwicki, J.E. and Strong, L.E. (2000) *Current Opinions in Chemical Biology*, **4**, 696–703.

Kim, Y.S., Gum, J.R. and Brockhausen, I. (1997) *Glycoconjugate Journal*, **13**, 693–707.

Kobata, A. (1993) *Account Chemistry Research*, **26**, 319–24.

Kuster, B., Hunter, A.P., Wheeler, S.F. *et al.* (1997) *Analytical Biochemistry*, **250**, 82–101.

Leppanen, A., White, S.P., Helin, J. *et al.* (2000) *Journal of Biological Chemistry*, **275**, 39569–78.

Lohmann, K.K. and von der Lieth, C-W. (2004) *Nucleic Acids Research*, **32**, W261–6.

Loris, R. (2002) *Biochimera Biophysica Acta*, **1572**, 198–208.

Lowe, J.D. and Marth, J.D. (2003) *Annual Reviews in Biochemistry*, **72**, 643–91.

Lütteke, T., Frank, M. and von der Lieth, C-W. (2004) *Carbohydrate Research*, **339**, 1015–20.

Lütteke, T., Frank, M. and von der Lieth, C-W. (2005) *Nucleic Acids Research*, **33**, D242–6.

MacBeath, G., Koehler, A.N. and Schreiber, S.L. (1999) *Journal American Chemistry Society*, **121**, 7967–8.

Mackiewicz, A. and Mackiewicz, K. (1995) *Glycoconugate Journal*, **12**, 241–7.

McNaught, A.D. (1997) *Carbohydrate Research*, **297**, 1–90.

Narimatsu, H. (2004) *Glycoconjugate Journal*, **21**, 17–24.

Park, K.-H., Robyt, J.F. and Choi, Y.-D. (1996) *Enzymes for Carbohydrate Engineering*, Elsevier, New York.

Plante, O.J., Palmacci, R. and Seeberger, P.H. (2001) *Science*, **291**, 1523–7.

PRUSINER, S.B. (1996) *Trends in Biochemical Sciences*, **21**, 482–7.

ROBYT, J.F. (1998) *Essentials of Carbohydrate Chemistry*, Springer-Verlag, New York.

RUDD, P.M., JOAO, H.C., COGHILL, E. *et al*. (1994) *Biochemistry*, **33**, 17–22.

RUDD, P.M., ELLIOTT, T., CRESSWELL, P. *et al*. (2001) *Science*, **291**, 2370–8.

SAXON, E. and BERTOZZI, C.R. (2000) *Science*, **287**, 2007–10.

STENUTZ, R., JANSSON, P-E., and WIDMALM, G. (1998) *Carbohydrate Research*, **306**, 11–17.

SUSUKI, S. and HONDA, S. (1998) *Electrophoresis*, **19**, 2539–60.

TAGLIARO, F., CRIVELLENTE, F., MANETTO, G. *et al*. (1998) *Electrophoresis*, **19**, 3033–9.

TAKETA, K. (1998) *Electrophoresis*, **19**, 2595–602.

TANIGUCHI, N., EKUNI, A., KO, J.H. *et al*. (2001) *Proteomics*, **1**, 239–47.

TOLBERT, T.J. and WONG, C-H. (2000) *Journal of the American Chemistry Society*, **122**, 5421–8.

VAN DAM, G., BOGITSH, B., VAN ZEYL, R. *et al*. (1996) *Journal of Parasitology*, **82**, 557–64

VAN DEN STEEN, P., RUDD, P.M., DWEK, R.A. and OPDENAKKER, G. (1998) *Critical Reviews in Biochemical Molecular Biology*, **33**, 151–208.

VARKI, A. (1997) *FASEB Journal*, **11**, 248–55.

VARKI, A., CUMMINGS, R., ESKO, J. *et al*. (1999) *Essentials of Glycobiology*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

WELLS, L., GAO, Y., MAHONEY, J.A. *et al*. (2002) *Journal of Biological Chemistry*, **277**, 1755–61.

WINCHESTER, B., CLAYTON, P., MIAN, N. *et al*. (1995) *Biochemistry Society Translations*, **23**, 185–8.

ZHANG, H., LI, X.L., MARTIN, D.B. and AEBERSOLD, R. (2003) *Nature Biotechnology*, **21**, 660–6.

World Wide Webs cited

CarbBank:	http://www.boc.chem.uu.nl/sugarbase/carbbank.html
Carbohydrate Structure Suite (CSS):	http://www.dkfz.de/spec/css/
CASPER:	http://www.casper.organ.su.se/casper/
CAZY:	http://afmb.cnr-mrs.fr/CAZY/
Consortium for Functional Glycomics:	http://www.functionalglycomics.org/
Distance Mapping:	http://www.glycosciences.de/distmap/
Dynamic Molecule:	http://www.mol-simulation.de/
ExPaSy tools:	http://www.expasy.org/tools
GlycoFragments:	http://www.glycosciences.de/GlycoFragments/fragment.php4
Glycosciences:	http://www.glycosciences.de/index.php
Glycosciences, Databases:	http://www.glucosciences.de/sweetdb/index.php
Glycosciences, Modeling:	http://www.glucosciences.de/modeling/index.php
Glycosciences, Tools:	http://www.glucosciences.de/tools/index.php
GlycoSearchMS:	http://www.dkfz.de/spec/glycosciences.de/sweetdb/ms/
GlycoWord:	http://www.gak.co.jp/FCCA/glycoword/wordE.html
IUBMB:	http://www.chem.qmw.ac.uk/iubmb
KEGG Carbohydrate Matcher (KCaM):	http://glycan.genome.ad.jp
Lectin DB:	http://cermav.cnrs.fr/
LinearCode for glycans:	http://www.glycomics.com/
LINUCS:	http://www.dkfz.de/spec/linucs
O-Glycobase:	http://www.cbs.dtu.dk/databases/OGLYCBASE/
PDB:	http://www.rcsb.org/pdb/
Pdb2linucs	http://www.dkfz.de/spec/pdb2linucs/
SweetDB:	http://www.glycosciences.de/sweetdb/index.php
Glycan structure databases:	Table 17.3
Glycan analysis servers:	Table 17.4
Utilities/tools for glycomics:	Table 17.5

CHAPTER 18

BIOMACROMOLECULAR EVOLUTION

18.1 VARIATION IN BIOMACROMOLECULAR SEQUENCES

18.1.1 Mutation as driving force of evolution

It is postulated that all extant organisms arose from a common ancestor that had already acquired all the basic common biochemical features. Although functional differences among organisms have arisen during evolution, the most basic biochemical processes have been conserved because any alternation to them would have been lethal. Comparison of DNA, RNA and protein sequences (biosequences) has become acceptable as the best means of reconstructing the process of evolution because the genetic information contained in biosequences is so vast (Higgs and Attwood, 2005). The usefulness of biosequences in reconstructing evolution is the realization that sequence similarity implies descent from a common ancestor (vertical descent), i.e. evolutionary divergence. Most frequently, changes result from point mutations such as insertion/deletion (InDel) and substitution. These events may accumulate through random drift and natural selection leading to a gradual divergence of initially identical gene copies. Where such divergence occurs, orthologs result between genes in different organisms in which divergent genes under selective pressure maintain the ancestral line. On the other hand, where such divergence occurs between duplicated genes in the same organism, paralogs result to reflect the adaptation of individual copies to separate but often related functions. Such paralogy can lead to protein families with numerous members in the same organisms, such as the human G-protein coupled receptors (Horn *et al.*, 2001). The diversity of paralogous families is further enhanced by recombination resulting in chimeric forms with novel properties. Among prokaryotes (particularly bacteria), genes have been passed between genomes known, as horizontal gene transfer (Garcia-Vallve *et al.*, 2000).

As sequence evolution is mainly divergent, it is relatively easy to reconstruct evolutionary trees from biosequence data. Nucleic acids and proteins that have evolved from a common ancestor are said to be homologous. That is, the sequences of homologous genes and proteins were identical at the time when they originated by replication of a single gene. The other explanation, other than divergence for similarities among biosequences, is convergence in which two or more unrelated sequences have become similar under the pressure of selection for similar functions. Convergence is a common evolutionary phenomenon at the level of protein three-dimensional structure and glycan structure.

Evolution occurs through genome variation followed by selection. Mutations are the driving force of evolution whereas natural selection modulates the rate of divergence. Mutation is the ultimate source of the genetic variation required for adaptation. Although the spontaneous mutation rate is generally low (~10^{-8} per nucleotide per generation in

---

Biomacromolecules, by C. Stan Tsai
Copyright © 2007 John Wiley & Sons, Inc.

humans corresponding to some 60 mutations on average in each individual). Genes undergo a continually stochastic process of mutation and every gene is subject to selective pressure. Any gene suffering a mutation that diminishes its chance of being passed on to the next generation is subject to negative selection and has a greater chance of disappearing from the population. For example, a tendency to change some parts of the genome such as those encoding the active sites and structural scaffolds of proteins, is more risky than a tendency to change others. Any gene with a mutation that gives it a greater than average chance of being passed on is subject to positive selection and has a better chance of remaining in the population. Mutations that affect a gene's ability to survive can be either deleterious or beneficial, but many mutations appear to have insignificant effects known as neutral mutation. Such mutation differs for each nucleotide in each gene, though neutral mutations will have to occur to every gene during its descent.

Optimal genetic variation results in part from the emergence of genes that increase diversity among a genome's descendants by a variety of mechanisms. These are called evolutionary genes. Similarly, selection would lead to the emergence of evolutionary information within a genome that controls the rate of evolution along the genome. Duplication of functional DNA followed by variation is a more efficient route to evolution of a new functional protein than is the random mutation of a noncoding stretch of DNA. In addition, duplication allows a genome to explore variation around a functional framework without losing the function of the original copy. The efficiency of duplication/variation is suggested by the emergence of large families of distinct genes that have related, conserved and useful structural scaffolds and/or active sites.

If a sufficient number of related sequences are available, the origin of any gene and protein can be determined, and the evolutionary history of its divergence can be inferred. Reconstructing an evolutionary pathway by comparing contemporary sequences is to some extent an educated guesswork because the sequences of the ancestral genes or proteins are usually not available. Most organisms are related at the molecular level. The more closely related they are evolutionarily, the more similar they are biochemically and genetically. Thus the more closely related the organisms, the more similar the sequences of their genes and proteins.

In comparing distantly related proteins, it is helpful to consider the chemical natures of the amino acid residues because analysis of closely related proteins demonstrate that chemically similar amino acids most often replace each other in substitution mutations. The divergence constraints on change in the amino acid sequences of related proteins derive from the necessity of those proteins to perform specific function vital to the survival of organisms. The gene sequences coding for closely related proteins also contribute to the constraints on divergence. Nucleotide changes at the third position of a codon that do not alter the specified amino acid occur more frequently than those that do. The sequences of introns, which have no known function, change more rapidly than do those exons. The upstream and downstream untranslated regions of genes also change more rapidly than the regions translated into polypeptide chains. The amino acid replacements that occur during protein divergence are nonrandom. The most prevalent replacements occur between amino acids with similar side chains, such as Gly/Ala. Ala/Ser, Ser/Thr, Val/Leu/Ile, Asp/Glu, Lys/Arg and Tyr/Phe.

The chemical bias by which certain amino acids change to varying extent and are replaced nonrandomly by certain other amino acids, does not arise from the nature of the genetic code or from the types of mutations that occur. The amino acid replacements that would be expected from random single nucleotide replacements are remarkably different from that actually observed among homologous proteins (Table 18.1). The numbers of observed replacements reflect the frequency with which each amino acid occurs in pro-

TABLE 18.1 Relative frequencies of amino acid replacement in a total of 1572 closely related proteins

	G	A	V	L	I	M	C	S	T	N	Q	D	E	K	R	H	F	Y	W	P
G		44	**38**				**16**	16				21	24		**43**			**13**		
A	58		41					35	39			23	26							36
V	10	37		48	18	12						**20**	**23**				**16**			
L	2	10	30		16	21		**13**			16				**22**	**11**	**41**		**9**	**26**
I		7	**66**	25		**25**		**15**	17	**18**							16			
M	1	3	8	21	6				11					**19**	7					
C	1	3	3		2			24							**10**		**12**	**15**	**16**	
S	**45**	**77**	4	3	2	2	12		41	15					**30**		**13**	**15**	**8**	24
T	5	59	**19**	5	13	3	1	**70**		17				24	**12**					**28**
N	**16**	**11**	1	4	4			**43**	17			19		49		13		**18**		
Q	3	9	3	8	1	2		5	4	5			23	24	13	26				15
D	16	15	2		1			10	6	**53**	8		49			**15**		**20**		
E	11	27	4	2	4	1		9	3	9	**42**	**83**		**30**						
K	6	6	2	4	4	9		**17**	20	32	15		10		20					
R	1	3	2	2	3	2	1	14	2	2	12	9		**48**		8			**13**	**20**
H	1	2	3	4			1	3	1	23	24	4	2	2	10			**13**		10
F	2	2	1	17	9	2		4	1	1					1	2		16		
Y		2	2	2	1		3	2	2	4			1	1		4	26			
W				1				2							3		1	1		
P	5	35	5	4	1		1	27	7	3	9	1	4	4	7	5	1			

Notes: 1. Amino acid replacements expected for random single-nucleotide mutations (upper right) compare with observed replacements (bottom left). The greatest discrepancies between the observed and random replacements are shown in boldface type.
2. Adapted from Dayhoff (1978).

teins (Table 18.2). The table also lists the normalized values for the relative mutabilities of amino acids. This bias in amino acid replacements presumably reflects the role of selection. Only those mutations that do not disrupt the function of the protein survive. As the consequence, chemically similar replacements are found the most frequently.

The evolutionary variations of a nucleic acid/protein also provide much information about the nucleic acid/protein itself. Evolutionary divergence into different species has resulted in many variants of the same functional nucleic acid/protein with different nucleotide/amino acid sequences. The differences and similarities of the biosequences of these variants reflect the constraints of structure and function for that nucleic acid/protein.

18.1.2 Evolutionary rate and role of selection

The rate of change of a gene is expressed as the number of single base-pair mutations per nucleotide position per unit of time. It is analogous to a mutation rate. Different genes/proteins have evolved at different rate (Table 18.3). If the times of divergence of the various species are known from other data, the rate at which each gene/protein has changed can be calculated.

It has been suggested somewhat controversially, that the rate of occurrence of mutations in a gene is constant in all species. Therefore the rate of evolutionary change of any gene or protein can serve as an evolutionary clock, with each gene/protein clock ticking at a particular rate in all species. If this is true, it has great implications for studying the mechanism of molecular evolution and evolutionary relationships among species (Wilson

TABLE 18.2 Frequencies of occurrence and relative variabilities of amino acids in proteins

Amino acid residues	Frequencies in proteins (%)	Variability/ Mutability
Asn	4.4	100
Ser	6.9	90
Asp	5.3	79
Glu	6.2	76
Ala	8.3	75
Ile	5.2	72
Thr	5.8	72
Met	2.4	70
Gln	4.0	69
Val	6.6	55
His	2.2	49
Arg	5.7	49
Lys	5.7	42
Pro	5.1	42
Gly	7.2	37
Phe	3.9	31
Tyr	3.2	31
Leu	9.0	30
Cys	1.7	15
Trp	1.3	13

Notes: 1. Frequencies of occurrence of amino acid residues in 1021 unrelated proteins, taken from McCalden and Argos (1988).
2. Variability expresses the number of times that a given amino acid residue has changed during the evolution of various homologous proteins divided by the number of times the residue occurred. The values have been normalized (highest value to 100), taken form Dayhoff (1978).

et al., 1977). The application of a constant evolutionary clock is that once the rate of change is established for a gene/protein, that rate can be used with the appropriate sequences to estimate the time of divergence of any pair of species. Rapidly changing genes/proteins are useful for studying closely related species and slowly changing genes/proteins reveal more distant evolutionary relationship. What is most surprising about the constant rate of evolutionary change of each gene/protein is that it remains constant per unit of time in species with greatly different generation times. The greatest exceptions to constant rates of divergence are those instances in which the function of the protein has changed.

Even if the total mutation rate is the same for all genes, the neutral mutation rate would differ for each gene because of the different fractions of mutations that are effectively neutral (mutation which does not significantly affect protein function and therefore has not been selected for or against). Every gene/protein would differ from every other in how much its amino acid sequence can vary without affecting its function. If the exact amino acid sequence is not critical for the function of a protein, a large fraction of its total mutations would be neutral, and the sequence of the protein would evolve rapidly. For proteins with very few acceptable amino acid replacements, their sequences would evolve at very slow rates. Generally, the degree of change in a protein sequence is found to be inversely proportional to the biological importance of each residue. The most variable residues are those that occur on the surface of a protein but are not functionally active.

TABLE 18.3 Evolutionary rates of some proteins

Protein	PAM/10^6 yrs	Protein	PAM/10^6 yrs
Histones:		Fibrous proteins:	
H4	0.0025	Collagen (α-1)	0.028
H3	0.0030	Crystallin (αA)	0.045
H2 (A, B)	0.017	Hormones:	
H1	0.12	Glucagon	0.023
		Corticotrpin	0.042
Enzymes, intercellular:			
Glutamate DH	0.018	Insulin	0.071
Triosephosphate DH	0.050	Thyrotripin β chain	0.11
TPI	0.053	Lipotropin β chain	0.13
Lactate DH (H_4)	0.053	Lutropin α chain	0.14
Lactate DH (M_4)	0.077	Proparathyrin	0.14
Carbonic anhydrase B	0.25	Porlactin	0.20
Carbonic anhydrase C	0.48	Growth hormone	0.25
		Lutropin β chain	0.33
Enzymes, secreted:			
Trypsinogen	0.17	Immunoglobulins:	
Lysozyme	0.40	γ (H) chains (C region)	0.59
Ribonuclease A	0.43	λ (L) chains (C region)	0.59
		λ (L) chains (V region)	1.25
Electron carriers:			
Cytochrome c	0.067	γ (H) chains (V region)	1.43
Cytochrome b_5	0.091	Other proteins:	
Plastocyanin	0.14	Parvalbumin	0.20
Ferredoxin	0.17	Albumin	0.33
		α-Lactalbumin	0.43
Oxygen-binding proteins:			
Myoglobin	0.17	Fibrinopeptide A	0.59
Hemoglobin α chain	0.27	Casein κ chain	0.71
Hemoglobin β chain	0.30	Figrinopeptide B	0.91

Notes: 1. One PAM (Point accepted mutation) is a unit of evolutionary divergence in which 1% of the amino acids have been changed.

2. Abbreviations used are: DH, dehydrogenase; TPI, triosephosphate isomerase.

3. Adapted from Wilson *et al*. (1977).

The most conserved amino acid residues are those that are most directly involved in the biological function of the protein. Similarly for DNA sequences, the noncoding regions of genes (introns), vary much more than the coding regions (exons). Within the coding regions, nucleotides that do not alter the amino acid sequence (degenerate codons) change more frequently than others.

The neutral mutation rate differs for each nucleotide in a gene and usually serves as a good indicator for the functional importance of encoded amino acid residue. For example, the C-peptide of proinsulin evolves at a rate (0.526 PAM/10^6 years) much more rapid than that of the A- and B-chains of insulin (0.071 PAM(point accepted mutation)/10^6 years) because the C-peptide appears to promote the protein folding and is removed proteolytically after the insulin has folded to its correct conformation. The greater divergent rate of the C-peptide than that of the A and B-chains reflects the fewer constraints on its amino acid sequence relative to the biologically functional chains of the hormone. The changes at the molecular level that have occurred during evolution are those that are least likely to have functional consequences. Thus the occurrence of nonfunctional changes is the result of the accumulation of neutral mutations. Natural selection at the molecular level appears primarily to be negative by weeding out the deleterious mutations that affect function. Although there are instances at the molecular level in which natural selection has had a positive effect in selecting for favorable mutations and functional differences, most evolutionary divergence of genes/proteins is probably of the neutral variety without significantly affecting their functions.

18.2 ELEMENT OF MOLECULAR PHYLOGENY

Systematics is the science of comparative biology. The primary goal of systematics is to describe taxic diversity and to reconstruct the hierarchy, or phylogenetic relationships, of those taxa. The grouping of organisms according to their natural relationships is known as taxonomy. The NCBI offers a number of different taxonomic resources, including the use of BLAST (www.ncbi.nim.nih.gov/entrez/query.fcgi?db=Taxonomy) to allow the researcher to construct evolutionary relationship diagrams, or trees for many different organisms:

- Phylogenetic inference is the process of developing hypotheses about the evolutionary relatedness of organisms based on their observable characteristics. Molecular biology has offered systematicists an almost endless array of characters in the forms of DNA, RNA and protein sequences with different structural/functional properties, mutational/selectional biases and evolutionary rates. The analysis/comparison of biosequences in phylogenetic inference is referred to as molecular phylogenetics (Page and Holmes, 1998; Nei and Kumar, 2000). The tremendous flexibility in their resolving power ensures the importance and widespread acceptance of biosequences in phylogenetic research. Phylogenetic analysis is the means of inferring or estimating evolutionary relationships. The physical events yielding a phylogeny happened in the past, and they can only be inferred or estimated. Phylogenetic analysis of biological sequences assumes that the observed differences between the sequences are the result of specific evolutionary processes. These assumptions are:
- The evolutionary divergences are strictly bifurcating, so that the observed data can be represented by a treelike phylogeny.
- The sequence is correct and originates from a specific source.
- If the sequences analyzed are homologous, they are descended from a shared ancestral sequence.
- Each multiple sequence included in a common analysis has a common phylogenetic history with the others.
- The sampling of taxa is adequate to resolve the problem of interest.
- The sequence variability in the sample contains a phylogenetic signal adequate to resolve the problem of interest.
- Each position in the sequence evolved homogeneously and independently.

The evolutionary history inferred from phylogenetic analysis is usually depicted as branching (treelike) diagrams. The correct evolutionary tree is a rooted tree that graphically depicts the cladistic relationships that exist among the operational taxonomic units

(OUTs) e.g. contemporary sequences. A tree consists of nodes connected by lines, and a rooted tree starts from the root, which is the node ancestral to all other nodes. Each ancestral node gives rise to two descendant nodes (in bifurcating trees). Terminal or exterior nodes having no further descendants correspond to OUTs. All nonroot, nonexterior nodes are interior nodes. They correspond to ancestral sequences and may be inferred from the contemporary sequences. The connecting lines between two adjacent nodes are branches of the tree. A branch represents the topological relationship between nodes. Branches are also divided into external and internal ones. An external branch connects an external node and an internal node, whereas an internal branch connects two internal nodes. The number of unmatched sites between the aligned sequences of the two adjacent nodes is the length of that branch. The number of these sequence changes counted over all branches constitutes the length of the tree. Trees can be made up of multigene families (gene trees) or a single gene from many taxa (species trees). The internal nodes correspond to gene duplication events in the first case and represent speciation events in the second case.

Calculation of tree length is simplified by removing the root from the tree. Such an unrooted tree will retain the interior nodes and the exterior nodes. A tree of N exterior nodes has N – 2 interior nodes and 2N – 3 links. Thus a tree of N OUTs can be converted to 2N – 3 dendrograms. The phylogenetic procedure can reconstruct ancestral sequences for each interior node of a tree but cannot determine which interior node or which pair of adjacent interior nodes is closer to the root. However, after the search for the optimal tree is terminated, we can use the full weight of evidence on the cladistic relationships of the OUTs to assign a root to the optimal tree or, at least, to some of its branches. In the out-group method of rooting, one (or more) sequence that is known to be an out-group to the N sequences of the unrooted tree is added. This converts the unrooted tree of N sequences to the recalculated rooted tree of N+1 sequences.

The number of possible tree topologies rapidly increases with an increase in the number (N) of OUTs. The general equation (Miyamoto and Cracroft, 1991) for the possible number of topologies for bifurcating unrooted trees (T_N) with n (≥3) OUTs (taxa) is given by

$$T_N = (2N - 5)!/[2^{N-3}(N - 3)!]$$

It is clear that a search for the true phylogenetic tree by phylogenetic analysis of a large number of sequences is computationally demanding and difficult.

A node and everything arising from a tree as a grouping is a ‘clade’ or a monophyletic group. A monophyletic group is a natural group such that all members are derived from a unique common ancestor and have inherited a set of unique common traits (characters) from it. A group excluding some of its descendents is a paraphyletic group. A mixed assembly of distantly related OUTs resembling one another or retaining similar primitive characteristics is polyphyletic:

Phylogenetics is the science of estimating the evolutionary past, in the case of molecular phylogeny, based on the comparison of DNA, RNA or protein sequences. It aims to reconstruct a phylogenetic tree as a statistical inference of a true phylogenetic tree, which is unknown. Various assumptions/approximations are made, a number of potential problems need to be considered:

- Scarcity/mixtures of the data can produce misleading results.
- Homologous sequences are commonly identified using an alignment algorithm. However, alignments of distantly related sequences may not be feasible and different alignment algorithms may produce variable results.

- A variety of methodological difficulties accompany different tree-making algorithms. Variation of mutational rate patterns among sites and lineages, multiple and retromutation events, functionally constrained sites and ancient evolutionary events make the estimates of distances uncertain. The assumptions and approximations in assessing evolutionary distances make estimates less reliable as sequence similarity diminishes.
- Losses in phylogenetic information may arise from mutation saturation, lateral gene transfer and a reconstruction artifact, long branch attraction phenomenon (Brown and Doolittle, 1997; Philippe and Laurent, 1998).
- Different phylogenetic reconstructions may result for the same set of organisms based on analysis of different proteins, genes or noncoding sequences.

18.3 PHYLOGENETIC ANALYSIS OF BIOSEQUENCES

18.3.1 General consideration

Phylogenetic analysis is any study in which biological variation is compared across samples. The basis of phylogenetic analysis arises from the fact that the objects of study are connected through historical relationships. Phylogenetic analysis of biosequences is a powerful tool for sorting and interpreting molecular data. It is possible to glean valuable information from a phylogenetic tree on the origin, evolution and possible function of genes and proteins they encode. Phylogenetic analysis has a wide range of applications such as reconstruction of the ancestral gene sequences from which extant genes are derived, classification/correlation of protein structures and functions, study of the origin and epidemiology of human diseases, inference to the evolution of ecological and behavioral traits through time, estimation of historical biogeographic relationships and exploration of the historical relationships across all of life. All phylogenetic analysis requires a criterion known as an optimality criterion for assessing how well the data fit candidate trees, a method for searching among possible solutions for the tree with the best fit to the data and a method for assessing confidence in the results. In molecular phylogenetics, the five basic steps in phylogenetic analysis of sequences (Hillis *et al.*, 1993) as organized in Figure 18.1 consist of:

1. Sequence alignment
2. Assessing phylogenetic signal
3. Choosing methods for phylogenetic analysis
4. Construction of optimal phylogenetic tree
5. Assessing phylogenetic reliability

18.3.2 Sequence data

The first step is building the sequence dataset (Durbin *et al.*, 1998), which generally involves searching and retrieving sequences from the public domains as referred to in subsection 14.2.1 (e.g. GenBank/EBI/DDBJ for nucleic acid sequences or PIR-PSD/Swiss-Prot/UniProt for protein sequences).

The sequences under study must be aligned so that positional homologues may be analyzed. Most methods of sequence alignment are designed for pair-wise comparison. Two approaches to sequence alignments are used, a global alignment (Needleman and

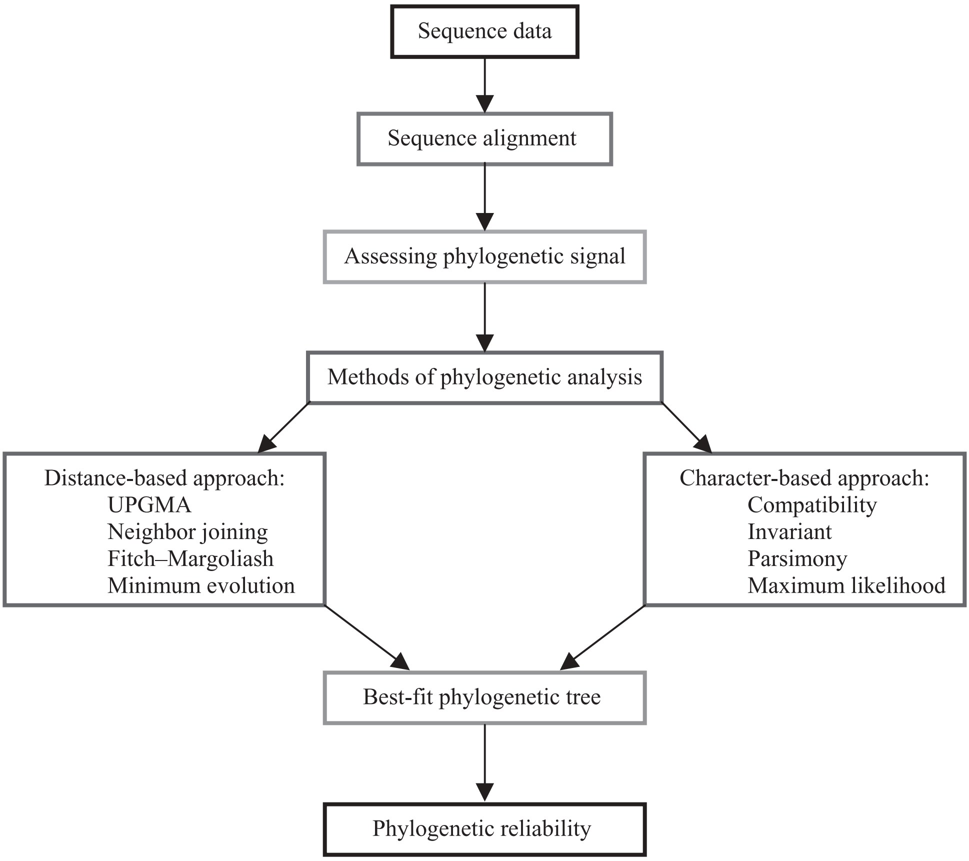

Sequence data
Sequence alignment
Assessing phylogenetic signal
Methods of phylogenetic analysis
Distance-based approach:
UPGMA
Neighbor joining
Fitch–Margoliash
Minimum evolution
Character-based approach:
Compatibility
Invariant
Parsimony
Maximum likelihood
Best-fit phylogenetic tree
Phylogenetic reliability

Figure 18.1 Flow chart for phylogenetic analysis of biosequences

Wunsch, 1970) and a local alignment (Smith and Waterman, 1981). The former, which compares similarity across the full stretch of sequences is the principal method of alignment for phylogenetic analysis whereas the latter, which searches for regions of similarity in parts of the sequences is better suited for searching databases (Chapter 14). Modifications of these approaches can also be used to align multiple sequences. Because the multiple alignment is inefficient with sequences if insertions/deletions (InDels) are common and substitution rates are high, most studies restrict comparisons to regions in which alignments are relatively obvious. Unless the number of taxa is few and InDels are uncommon, it is not feasible to ensure that the optimum alignment has been achieved. For this reason, regions of questionable alignments are often removed from consideration prior to phylogenetic analysis. In general, substitutions are more frequent between nucleotides/amino acids that are biochemically similar (Table 14.2). In the case of DNA, the transition between purine→purine and pyrimidine→pyrimidine are usually more frequent than the transversion between purine→pyrimidine and pyrimidine→purine. Such biases will affect the estimated divergence between two sequences. Specification of relative rates of substitution among particular residues usually takes the form of a square matrix, the number of rows/columns is 4 for DNA and 20 for proteins and 61 for codons. The diagonal elements represent the cost of having the same nucleotide/amino acid in different sequences. The off diagonal elements of the matrix correspond to the relative costs of going from one nucleotide/amino acid to another.

The phylogenetic analysis of biosequences seeks to maximize a similarity by means of the global pair-wise alignment. Operationally a series of pair-wise alignments are performed and these subalignments amalgamated into a multiple alignment. The intermediate pair-wise alignments are added together following a tree-like pattern. The order for carrying out the alignment is determined by a 'guide tree' in which each node of the tree represents a separate pair-wise alignment. The phylogeny resulting from a multiple alignment analysis is dependent on the order in which the sequences are accreted. Several methods centered on two areas are in use to progressively align biosequences by pair-wise accretion, i.e.:

1. techniques of establishing an alignment topology (guide tree); and
2. methods for combining the aligned sequence positions at the nodes to create the complete multiple alignment (Phillips *et al.*, 2000).

A clustal program such as ClustalW (Higgins *et al.*, 1996), which aligns sequences according to an explicitly phylogenetic criterion, is the most commonly used program for the multiple alignment of biochemical sequences (ftp://ft-igbmc.u-strasbg.fr/pub/). The CLUSTAL aligns the most similar sequence first. A consensus sequence is substituted for the sequence pair. Consensus sequences incorporate only bases/amino acids present in all sequences or use partial (75%) consensus. It progressively aligns the next most similar sequence to the consensus of the growing cluster or the next two most similar sequences to each other. There are other methods based on sequence similarity, such as BioEdit (http://www.mbio.ncsu.edu/BioEdit/bioedit.html), TREEALIGN (ftp://ftp.ebi.ac.uk/pub/sofware/unix/treealign.tar.Z) and MALIGN (ftp://ftp.amnh.org/pub/people/wheeler/malign/). These methods differ in how the phylogenetic tree is determined/selected and how the tree interacts with the actual alignment.

For most sequence data, some positions are highly conserved, whereas other positions are randomized with respect to phylogenetic history. Thus assessment of the phylogenetic signal is often required. In general, pair-wise comparisons of the sequences are performed to evaluate the potential phylogenetic importance of the data. For example, the transition/transversion ratios for sequence pairs of DNA can be compared to those expected for the sequences at equilibrium, given the observed base compositions. DNA sequences that are largely free of homoplasy (convergence, parallelism and reversal) will have transition/transversion ratios greater than those for sequences that are saturated by change, but similar to those observed for closely related taxa, which remain highly structured. Similarly, pair-wise divergences have been used to assess the potential phylogenetic value of DNA sequences, by plotting percent divergence against time. In such plots, regions of sequences or classes of character state transformations that are saturated by change do not show a significant positive relationship with time. Both the transition/transversion ratio and sequence divergence are influenced by homoplasy, thus both can provide insights into the potential phylogenetic value of sequence data. Another way of detecting the presence of a phylogenetic signal in a given data set is to examine the shape of the tree-length distribution (Fitch, 1979). Data sets with skewed tree-length distributions are likely to be phylogenetically informative. The data sets that produce significantly skewed tree-length distributions are also likely to produce the correct tree topology in phylogenetic analysis.

The following considerations may guide a choice of methods for phylogenetic analysis:

- assumptions about evolution;
- parameters of sequence evolution;
- the primary goal of analysis;

- size of the data set; and
- the limitation of computer time.

An important consideration in building molecular trees from protein encoding genes is whether to analyze biosequences at the DNA or protein level. Generally DNA sequences are used for closely related sequences because there are more informative changes at the DNA level while protein sequences can be used for more distance relationships since amino acid sequences hold more information.

Numerous methods are available and each makes different assumptions about the molecular evolutionary process. It is unrealistic to assume that any one method can solve all problems, given the complexity of genomes/proteomes and their evolution. Phylogenetic analyses of sequences can be performed by making pair-wise comparisons of whole biosequences (i.e. distance-based approach) or by analyzing discrete characters such as aligned nucleotides and amino acids of the sequences (i.e. character-based approach). Deciding whether to use a distance-based or a character-based method depends on the assumptions and the goals of the study.

18.3.3 Phylogenetic method: Distance-based approaches

Distance-based methods use the amount of the distance (dissimilarity) between two aligned sequences to derive phylogenetic trees. A distance method would reconstruct the true tree if all genetic divergence events were accurately recorded in the sequence. Distance matrix methods simply count the number of differences between two sequences. This number is referred to as the evolutionary distance, and its exact size depends on the evolutionary model used. The actual tree is then computed from the matrix of distance values by running a clustering algorithm that starts with the most similar sequences or by trying to minimize the total branch length of the tree. Both the type and the position of a mutation have been incorporated into the parameters of divergence values. The simplest of the divergence measures is the two-parameter model of Kimura (Kimura, 1980). This model is designed to estimate divergence as well as the more complicated algorithms, over a broad range of divergences, without the need for additional specifics about the evolutionary process.

Methods for clustering distance data can be divided into those that provide single topology by following a specific series of steps (single-tree algorithms) and those that use optimality criterion to compare alternative topologies and to select the tree (multiple-tree algorithms). The representative single-tree algorithms are:

- the unweighed pair group method using arithmetic means (UPGMA); and
- the neighbor joining method (NJ).

The pair-wise clustering procedure used for UPGMA is intuitive. Each sequence is assigned to its own cluster to start a branch of the tree. The two clusters that are closest together in terms of whatever distance measure has been chosen are merged into a single cluster. A branch point (or node) is defined that connects the two branches. The node is placed to reflect the distance between the two leaves (sequences) that have been joined. The process is repeated iteratively, until there are only two clusters left. When they are joined, the root of the tree is defined. The branch lengths in a tree constructed using this process theoretically reflects evolutionary time. The NJ (Saitou and Nei, 1987) algorithm searches not just for minimum pair-wise distances according to the distance metric, but also for sets of neighbors that minimize the total length of the tree. It has become a popular approach for analyzing sequence distances because of its ability to handle unequal rates, its connection to minimum-length trees and its ease of calculation with regard to both tree topology and branch length.

Several criteria of optimality are used for building distance trees. Each approach permits unequal rates and assumes additivity. The Fitch–Margoliash method (Fitch and Margoliash, 1967) minimizes the deviation between the observed pair-wise distances and the path length distances for all pairs of taxa on a tree. The best phylogenetic tree maximizes the fit of observed (original) versus tree-derived (patristic) distances, as measured by % standard deviation. Branch lengths are determined by linear algebraic calculations of observed distances among three different taxa interconnected by a common node. Minimum evolution (Saitou and Imanishi, 1989) aims to find the shortest tree that is consistent with the path lengths measured in a manner similar to that of the Fitch–Margoliash approach without using all possible pair-wise distances and all possible associated tree path lengths. Rather it chooses the location of internal tree nodes based on the distance to external nodes, and then optimizes the internal branch length according to the minimum measured error between these observed points. The phylogenetic tree is the one with the minimal total overall length.

18.3.4 Phylogenetic method: Character-based approaches

The individual aligned sites of biosequences are equivalent to characters. The actual nucleotide or amino acid occupying a site is the character state. The character-based approaches treat each substitution separately rather than reducing all of the individual variation to a single divergence value. The relationships among organisms by the distribution of observed mutations are determined by counting each mutation event. These methods are preferred for studying character evolution, for combining multiple data sets and for inferring ancestral genotypes. All sequence information is retained through the analyses. No information is lost in the conversion to distances. The approaches include parsimony, maximum likelihood and method of invariant.

The principle of parsimony searches for a tree that requires the smallest number of changes to explain the differences observed among the taxa under study (Czelusnlak *et al.*, 1990). The method minimizes the number of evolutionary events required to explain the original data. Parsimony searches among the set of possible trees to find the one requiring the least number of nucleotide/amino acid substitutions to explain the observed differences between sequences. The topology that is maximally parsimonious is that for which the total number of inferred changes at all the informative sites is minimized. In practical terms, the parsimony tree is the shortest and the one with the fewest changes (least homoplasy). Parsimony remains the most popular character-based approach for sequence data due to its logical simplicity, its ease of interpretation, its prediction of both ancestral character states and amount of change along branches, the availability of efficient programs for its implementation and its flexibility of conducting character analyses.

The maximum likelihood method (Felsenstein, 1981) calculates the probability of a data set, given the particular model of evolutionary change and specific topology. The method searches for the optimal choice by assigning probabilities to every possible evolutionary change at informative sites, and by maximizing the total probability of the tree. The likelihood of changes in the data can be determined by considering each site separately with respect to a particular topology and model of molecular evolution. Therefore the maximum likelihood method depends largely on the model chosen and on how well it reflects the evolutionary properties of the biomacromolecule being studied. In practice, the maximum likelihood is derived for each base position in an alignment. An individual likelihood is calculated in terms of the probability that the pattern of variation produced at a site by a particular substitution process with reference to the overall observed base frequencies. The likelihood becomes the sum of the probabilities of each possible reconstruction of substitutions under a particular substitution process. The likelihoods for all the sites are multiplied to give an overall likelihood of the tree. Because of questions about the accuracy of the models, coupled with the computational complexities of the approach, maximum likelihood methods have not received wide attention.

18.3.5 Construction of phylogenetic tree

The best tree under the selected optimality criterion must be estimated (Saitou, 1996) once a method and appropriate software have been selected. The number of distinct tree topologies is very large, even for a modest number of taxa (e.g. over 2.8×10^{74} distinct, labeled, bifurcating trees for 50 taxa). For relatively few taxa (up to as many as 20 or 30) it is possible to use exact algorithms (algorithmic approach) that will find the optimal tree. For greater numbers of taxa, we must rely on heuristic algorithms that approximate the exact solutions but may not give the optimal solution under all conditions. The exact algorithm searches exhaustively through all possible tree topologies for the best solution(s). This method is computationally simple for 9 or fewer taxa and becomes impractical for 13 or more taxa (≥13 749 310 575 trees). An alternative exact algorithm known as the branch-and-bound algorithm (Hendy and Penny, 1982) can be applied to speed up the analysis of a data set consisting of less than 10–11 taxa. If the exact algorithms are not feasible for a given data set, various heuristic approaches can be attempted. Heuristic procedures (Stagle, 1971) are a computer simulation and programming philosophy suited to finding a problem solution exploiting any empirical trial, strategy or shortcut by which the computer acquires knowledge of the structure of the problem space beyond its pure abstract definition. The heuristic methodology does not, however, guarantee (in contrast to the algorithmic approach) the discovery of all possible goals in an absolute sense. Most heuristic techniques start by finding a reasonably good estimate of optimal tree(s) and then attempting to find a better solution by examining structurally related trees. The initial tree is usually found by a stepwise addition of taxa sequentially to a tree at the optimal place in the growing tree. Different representations of phylogenetic trees are illustrated in Figure 18.2. The subsequent improvement is achieved by examining related topologies by a series of procedures known as branch swapping. All involve rearranging branches of the initial tree to search for a shorter alternative.

Often branch lengths are drawn to scale in molecular phylogenetic trees, i.e. proportional to the amount of evolution estimated to have occurred along them. This gives a general impression of relative rates of changes across a tree. Bootstrap values are displayed as percentages. This makes the tree easier to read and to compare with other trees. By convention, only bootstrap values of 50% or higher are reported. Meaningful names for OUTs should be given and annotations to the tree by indicating important groups are desirable.

18.3.6 Assessment

When large amounts of homoplasy exist among distantly related branches, we can avoid the abundant homoplasy while recognizing the phylogenetic signal by relying on a few specific patterns of nucleotide variation that represent the most conservative changes (Lake, 1987). For example, three possible topologies for four taxa called operator invariants are calculated. These calculations are based on the variable positions with two purines and two pyrimidines. Zero-value invariants represent cases in which random multiple mutation events have canceled each other out, and as such a chi-squared (χ^2) or binomial test is used to identify the correct topology as the one with an invariant significantly greater than zero. There are 36 patterns of transitions/transversions for four taxa with three possible topologies. Different methods of phylogenetic inference rely on various combinations of these components, with 12 used to calculate the three operator invariants of evolutionary parsimony. When more than four taxa are considered, all possible groups of four are typically analyzed and a composite tree constructed from the individual results.

To assess confidence of phylogenetic results, it is generally recommended that subsets of taxa be examined from within the more complex topologies by focusing on specific major questions targeted before the analysis. Alternatively we could limit the comparisons to just those topologies considered plausible for biological reasons. The reliability in phylogenetic results can be assessed by analytical methods or resampling methods. Analytical techniques (Felsenstein, 1988; Lake, 1987) for testing phylogenetic reliability operate by comparing the support for one tree to that for another under the assumption of randomly distributed data. Resampling techniques estimate the reliability of a phylogenetic result by bootstrapping or jackknifing the characters of the original data set. The bootstrapping approach (Felsenstein, 1985; Hillis and Bull, 1993) creates a new data set of the original size by sampling the available characters with replacement. Whereas the jackknife approach (Penny and Hendy, 1986) randomly drops one or more data points or taxa at a time creating smaller data sets by sampling without replacement. The ultimate criterion for determining phylogenetic reliability rests on tests of congruence among independent data sets representing both molecular and nonmolecular information (Hillis, 1987). Different character types and data sets are unlikely to be exposed to the same evolutionary biases, and as such, congruent results supported by each are more likely to reflect convergence on to the single, correct tree. Therefore congruence analysis provides an important mechanism with which to evaluate the reliability of different methods for constructing phylogenetic trees. Given a common data set, those approaches leading to the congruent result should be preferred over those that do not. Studies of congruence can also provide insights into the limitations and assumption of different tree-constructing algorithms. Congruence analysis further permits an evaluation of the reliability of different weighting schemes of character transformations used in a phylogenetic analysis.

18.4 APPLICATION OF SEQUENCE ANALYSES IN PHYLOGENETIC INFERENCE

18.4.1 Phylogenetic analysis software

The catalog of available software programs for phylogenetic analysis is maintained at Phylogeny Programs of PHYLIP at http://evolution.genetics.washington.edu/phylip/software.html. Some of software programs for PC Windows are listed in Table 18.4.

18.4.2 Phylogenetic analysis with PHYLIP

Phylogenetic Inference Package (PHYLIP) is a software package comprising about 30 command-line programs that cover almost any aspect of phylogenetic analysis (Felsenstein, 1985; Felsenstein, 1996). The software (either PC Window or Apple version) can be downloaded from the PHYLIP web site and used locally. The package consists of a diverse collection of programs, including routines for calculating estimates of divergence and programs for both distance-based and character-based phylogenetic analyses.

TABLE 18.4 Some software programs for phylogenetic analysis of biosequences

Software	Program/method	Resource
Arlequin	DC, DM	http://acasun1.unige.ch/arlequin
Bionumerics	G: DC, DM, ML, P	http://www.aplied-maths.com/bn/bn.htm
COMPONENT	CT	http://taxonomy.zoology.gla.ac.uk/rod/cplite/guide.htm
DAMBE	G: DC, DM, ML, P	http://aix1.uottaw.ca/~xxia/software/software.htm
DnaSP	DC	http://www.ub.es/dnasp
EDIBLE	ML	http://ebi.ac.uk/goldman/info/edible.html
GDA	DM	http://hydrodictyon.eeb.uconn.edu/people/plewis/software.php
GelCompr II	DC, DM, P	http://www.aplied-maths.com/gc/gc.htm
GeneTree	P	http://taxonomy.zoology.gla.ac.uk/rod/genetree/genetree.htm
HY-PHY	ML	http://www.hyphy.org
MEGA	G: CT, DC, DM, ML, P	http://www.megasoftware.net
Mesquite	G: CT, DM, ML, P	http://mesquiteproject.org
Network	P	http://www.fluxus-engineering.com/sharenet.htm
Nona	P	http://www.cladistics.com/aboutNona.htm
PAL	G: DC, DM, ML, P	http://www.pal-project.org
PAML	DC, ML	http://atgc.lirmm.fr/phyml/
PAUP*	G: CT, DC, DM, I, ML, P	http://paup.csit.fsu.edu/
PHYLIP	G: CT, DC, DM, I, ML, P	http://evolution.genetics.washington.edu/phylip.html.
Phylo_win	G: DC, DM, ML, P	http://www.tmk.com/ftp/vms-freeware/methog/
Population	DC, DM,	http://www.cnrs-gif.fr/pge/bioinfo/populations/
SeqPup	DM, ML	http://iubio.bio.indiana.edu/soft/molbio/seqpup
SplitsTree	DM, ML	http://www-ab.informatik.uni-tuebingen.de/software/splits/
TreeCons	ML	http://bioinformatics.psb.ugent.be/software-details.php?id=3
TreeExoplorer	TM	http://evolgen.biol.metro-u.ac.jp/TE/TE-main.html
TreeJuxtaposer	TM	http://olduvai.sourceforge.net/tj/index.html
TREEMAP	CT	http://taxonomy.zoology.gla.ac.uk/rod/treemap.html
TreeView	TM	http://taxonomy.zoology.gla.ac.uk/rod/treeview.html
T-REX	DC, DM	http://www.labunix.uqam.ca/~makarenv/trex.html
Vanilla	DC, DM, ML	http://www.stat.uni-muenchen.de/~strimmer/pal-project/vanilla/
Winboot	DM	http://irri.org/science/software/winboot.asp

Notes: 1. Taken from Phylogeny Programs (http://evolution.genetics.washington.edu/phylip/software.html) for PC-Windows.
2. Abbreviations used: CT, consensus trees/supertrees; DC, distance calculation; DM, distance matrix; G, general; I, invariant method; ML, maximum likelihood; P, parsimony, TM, tree manipulation.

Table 18.5 lists some of the executable files (.exe). The general user's manual can be found in main.doc and detailed user's instruction to each executable file is in the separate documentation (.doc) file. The sequence data in an input file, called infile, must be in PHYLIP format, which can be obtained by the use of format converter, ReadSeq at http://dot.imgen.bcm.tmc.edu:9331/seq-util/Options/readseq.html.

With PHYLIP, DnaDist computes a distance matrix from nucleotide sequences. Phylogenetic trees are generated by any one of the available tools utilizing the distance matrix programs (Neighbor, Fitch or Kitsch). DnaDist allows the user to choose between three models of nucleotide substitution.

1. The Kimura two-parameter model allows the user to weigh transversions more heavily than transitions. ProtDist is a program that computes a distance matrix for an alignment of protein sequences. It allows the user to choose between one of three evolutionary models of amino acid replacements. The simplest, fastest model assumes that each amino acid has an equal chance of turning into one of the other 19 amino acids.

TABLE 18.5 Phylogenetic methods available from PHYLIP

Phylogenetic methods	Exe. program	Sequences	Description
Distance-based:			
Distance measure	DnaDist	DNA	Compute distances to distance matrix
Distance measure	ProtDist	Protein	Compute distances to distance matrix
Fitch-Margoliash (FM)	Fitch	DNA/Protein	Fitch-Margoliash method of analysis
FM with molecular cloak	Kitsch	DNA/Protein	FM method with molecular cloak
Neighbor joining	Neighbor	DNA/Protein	Neighbor joining analysis
Character-based:			
Compatibility	DnaComp	DNA	Search with compatibility criterion
Invariant	DnaInvar	DNA	Lake's phylogenetic invariant
Parsimony	DnaPars	DNA	Parsimony method
Parsimony	ProtPars	Protein	Parsimony method
Parsimony/compatibility	DnaMove	DNA	Interactive parsimony or compatibility
Maximum likelihood	DnaML	DNA	Max. likelihood without molecular cloak
Maximum likelihood	DnaMLk	DNA	Max. likelihood with molecular cloak
Search/reliability:			
Exact search	DnaPenny	DNA	Branch-and-bound search
Exact search	Penny	DNA/Protein	Branch-and-bound search 10–11 taxa or less
Resampleing	SeqBoot	DNA/Protein	Multiple data sets from bootstrap resampling
Statistic	Contrast	DNA/Protein	Independent contrast for multivariate statistics
Phylogenetic tree:			
Drawing	DrawGram	DNA/Protein	Plot rooted tree
Drawing	DrawTree	DNA/Protein	Plot unrooted tree
Consensus tree	Consensus	DNA/Protein	Compute consensus tree by majority-rule
Editing	ReTree	DNA/Protein	Reroot, flip branches and renaming

2. The second is a category model in which the amino acids are distributed among different groups and transitions are evaluated differently, depending on whether the change would result in an amino acid switching in the group.
3. The third (default) method uses a table of empirically observed transitions between amino acids (the Dayhoff PAM 001 matrix). The character-based analysis of sequence data can be initiated via the appropriate executable file (e.g. DnaPars, DnaML or ProtPars). PHYLIP comprises DnaPars and DnaML to estimate phylogenetic relationships by the parsimony method and the maximum likelihood methods from nucleotide sequences respectively. ProtPars is the parsimony program for protein sequences.

The acceptance of the input sequences is indicated by a return of the option menu. The program automatically searches for infile and writes the distance matrix into outfile. Copy/save outfile as otherfile for later reference and rename outfile to infile to be analyzed by one of the distance matrix programs (e.g. Neighbor, Fitch or Kitsch). The analytical results are recorded in outfile and treefile.

The treefile can be further processed with TreeView (Page, 1996), which allows the user to manipulate the tree and save the file in commonly used graphic formats. The tree-drawing software, TreeView can be downloaded from http://taxanomy.zoology.gla.ac.uk/rod/treeview.html, by choosing the tree view to launch one of the tree programs (radial, slanted cladogram or rectangular cladogram as exemplified in Figure 18.2) and by saving the graphic file as treename.wmf.

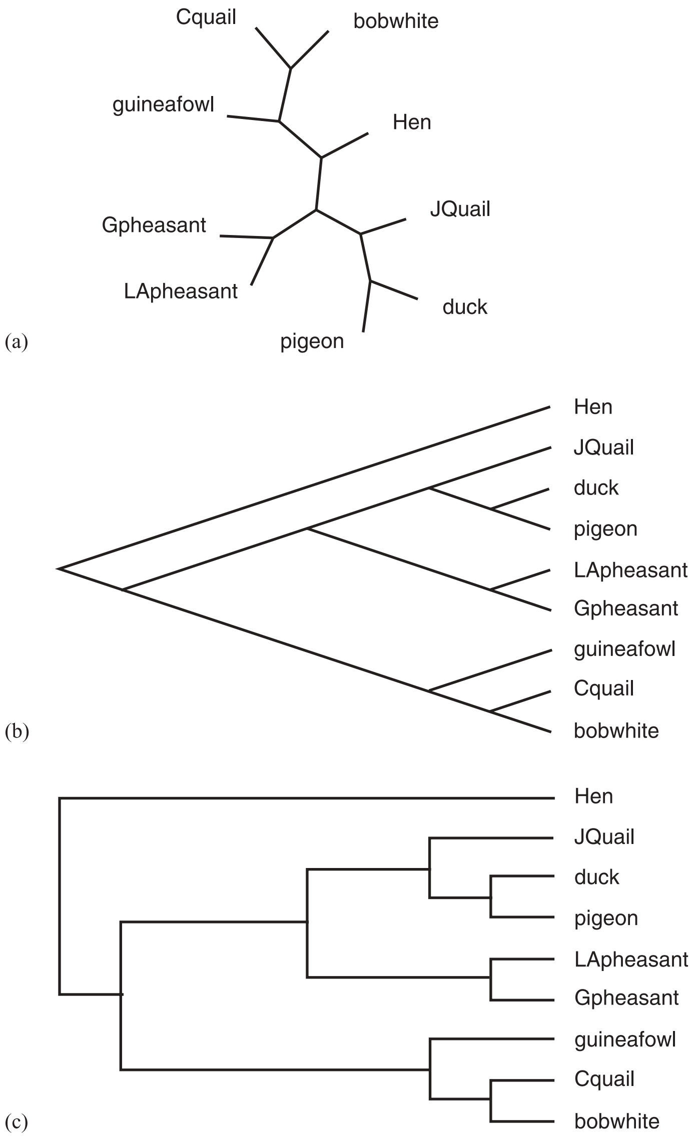

Cquail
bobwhite
guineafowl
Hen
JQuail
Gpheasant
LApheasant
duck
pigeon
(a)
Hen
JQuail
duck
pigeon
LApheasant
Gpheasant
guineafowl
Cquail
bobwhite
(b)
Hen
JQuail
duck
pigeon
LApheasant
Gpheasant
guineafowl
Cquail
bobwhite
(c)

Figure 18.2 Phylogenetic tree representations. Phylogenetic analysis (parsimony) of lysozyme sequences is conducted with Phylip. Phylogenetic trees (constructed with TreeView) are represented as radial (A), slanted cladogram (B) or rectangular cladogram (C)

To search for the consensus tree by bootstrapping techniques, Seqboot accepts an input from the infile and multiplies it a user-specified number of times (enter odd number n to yield a total of n+1 data sets). The resulting outfile, after renaming to infile, is subjected to either distance-based or character-based analysis. Copy/save treefile to otherfile (for later reference) and rename treefile to infile. The resulting trees are reduced to the single tree by the use of the Consensus program, which returns the consensus tree with the bootstrap values as numbers on the branches. The topology of the consensus tree of the outfile can be viewed with any text editor.

18.4.3 Phylogenetic analysis online

Clustal is the common program for executing multiple sequence alignment (Higgins *et al.*, 1996). After the alignment, the program constructs neighbor joining trees with bootstrapping. ClustalW can be accessed at EBI (http://www.ebi.ac.uk/), BCM (http://dot.imgen.bcm.tme.edu..9331/multi-align/multi-align.html) and DDBJ (http://www.ddbj.nig.ac.jp/). The tree file can be requested in PHYLIP format, which can be saved as aligname.ph and displayed with TreeView. The output alignment file can be saved as aligname.aln, which can be further analyzed using WebPHYLIP at http://sdmc.krdl.org.sg:8080/~lxzhang/phylip/welcome.html

The home page of WebPHYLIP consists of three windows. The left window is the command window for selecting and issuing the execution (Run) of the analysis/operation. The analysis results appear in the upper window. The lower window is the query window with options, a query box for pasting the query sequences (in PHYLIP format), and Submit/Clear buttons. To perform phylogenetic analysis, select DNA/Protein for Phylogeny method in the left window to open the available analytical methods (parsimony, parsimony + branch & bound, compatibility, maximum likelihood, and maximum likelihood with molecular clock for DNA whereas only parsimony for protein). Click Run (under the desired method) to open the query window with options. Select the appropriate Input type (either interleaved or sequential), paste the query sequences and click the Submit button. The analytical results are displayed in the upper window.

Select Do Consensus and click Run of Consensus tree (left window) to open the request form (lower window). Choose Yes to ‘Use tree file from last stage?’ and click the Submit button to display the consensus tree (upper window) and save the tree file. Select Draw trees and click Run of cladograms/phenograms/phylogenies. Choose Yes to ‘Use tree file from last stage?’ and click the Submit button to save the drawing treename.ps. Various phylogeny databases and utilities are listed in Table 18.6.

18.5 EVOLUTION OF BIOSEQUENCES

18.5.1 Evolution of nucleic acid sequence

Phylogenetic analyses often examine whether nucleotide substitutions are synonymous (not altering encoded amino acid) or nonsynonymous, and trace the history of gene-

TABLE 18.6 Phylogenetic databases and utilities

Web	URL	Description
COG DB	http://www.ncbi.nlm.nih.gov/COG	Clusters of orthologous proteins
Consensus alignment	http://structure.bu.edu/cgi-bin/consensus/consensus.cgi	Consensus align and modeling
CVTree	http://cvtree.cbi.pku.edu.cn	Phyl tree reconstruction
DED	http://warta.bio.psu.edu/DED/	
FootPrinter	http://bio.cs.washington.edu/software.html	Phyl footprinting software
HGMP	http://www.hgmp.mrc.ac.uk/	Phylogenetic linkage analysis
NCBI Taxonomy	http://www.ncbi.nlm.nih.gov/	All organism on GenBank
NEWT	http://www.ebi.ac.uk/newt/	Taxonomy portal
PALI	http://pauling.mbu.iisc.ernet.in/~pali	Phylogeny of protein structures
TreeDase	http://herbaria.harvard.edu/treebase	Phylogenetic trees
Tree of Life	http://phylogeny.arizona.edu/tree/phylogeny.html	Phylogeny and biodiversity info.
WebPHYLIP	http://sdmc.krdl.org.sg:8080/~lxzhang/phylip/	Phyl analysis of biosequences

Note: Abbreviation used: phyl, phylogenetic.

duplication events. The human Diaspora has been modeled by reconstruction phylogeny from mitochondrial sequences derived from people of different ethnic origins (Huelsenbeck and Imennov, 2002). Phylogenetic modeling has been used for projects such as predicting which strain of influenza is likely to hit next (Fitch *et al.*, 2000) or whether chimps are the origin of HIV (Paraskevis *et al.*, 2003). The basic purpose is to examine DNA and protein sequences in a comparative context. It is useful in classification at the molecular level, assessing biodiversity and timing major diversification events.

The universal tree of life represents a hierarchical phylogenetic classification of the living organisms based on comparative analysis of biosequences encoding rRNA and several proteins (Doolittle, 1999). The two major problems with inferring phylogenies from biosequences are:

1. the variation of similarity, and
2. different rates of evolution along different branches of the evolutionary tree.

However, phylogenies based on biosequences will remain indispensable. The choice of small subunit rRNA for the analysis/construction of universal tree derives from:

- It has massive database of sequence information
- It is abundant because it is coded for in organelle as well as nucleus.
- It has slow- and fast-evolving portions
- It has a universally conserved structure
- It is ancient and is involved in essential cell function
- It interacts with many other co-evolved cellular RNAs and proteins.

The universal tree based on small subunit rRNA sequences divides organisms into three domains, namely Archaea, Bacteria and Eukarya and the branching pattern of hundreds of subordinate taxa and is rooted in the prokaryotic bacterial domain (Woese, 1998). However, phylogenetic analysis of secondary structures of the small and large rRNA subunits indicates a reconstructed universal tree that branches in three monophyletic groups corresponding to Eucarya, Archaea and Bacteria and is rooted in the eukaryotic branch (Caetano-Anollé, 2002).

Certain noncoding repetitive sequences in genomes are useful for phylogenetic studies. Short interspersed nuclear element (SINE) and long interspersed nuclear element (LINE) are repetitive noncoding sequences that form large fractions of eukaryotic genomes (e.g. at least 30% of human chromosomal DNA). SINEs are tRNA-derived retroposons and are distinguished from LINEs by their large copy number (over 10^6 copies per haploid genome), relatively short length (70–500 bp for SINE versus up to 7000 bp for LINE), and inability to encode for enzymes (e.g. reverse transcriptase for their own amplification). Most SINEs contain a 5′ region homologous to tRNA, a tRNA-unrelated central region and a 3′ AT-rich region. SINE moves by a copy-and-paste mechanism known as retrotransposition, which is mediated by RNA. When inserted into functional regions of a genome, SINE can serve as insertable mutagens, alter gene expression or promote gene conversions and homologous recombination events. However, most SINEs insert innocuously into nonfunctional regions and can provide an excellent record of biological history that is largely free from character reversals and parallel evolution. Some features of SINE useful for the phylogenetic inference of species (Shedlock and Okada, 2000; Shedlock *et al.*, 2004) for example are:

- SINE is commonly found dispersed throughout the genome and can be primed for retrotransposition at nicked sites. Furthermore, horizontal transfer of SINE is severely restricted by their nonautonomous amplification.

- Copies of the same SINE shared by two different taxa are derived from the same initial insertion event in the germ line of a common ancestor; they define monophyletic groups or clades. Thus the presence of a SINE at any particular position is a property that entails uncomplicated and variable measure of common ancestry.
- SINE is inserted at random in the noncoding region of a genome. Therefore appearance of similar SINEs at the same locus in two species implies that the species share a common ancestor in which the insertion event occurred because there is no selection for the site of insertion.
- SINE insertion appears to be irreversible. If two species share a SINE at a common locus, absence of this SINE in a third species implies that the first two species must be more closely related to each other than either is to the third.
- SINE shows relationships, i.e. they imply which species came first. The last common ancestor of a species containing a common SINE must have come after the last common ancestor linking these species and another that lacks this SINE. Thus SINE insertion events allow the establishment of tree topologies but are unreliable for calculating relative branch length without reference to additional information.
- SINE flanking sequences provide valuable information on phylogenetic timing. The amount of divergence between SINE flanking sequences analyzed for clade will be proportional to the time that has elapsed since a diagnostic SINE locus originally appeared in the genome of a common ancestor. Because each SINE insertion represents an irreversible evolutionary event at time zero for a molecular clock corresponding to nucleotide change in its flanking sequences. Furthermore, as the majority of SINE insertions are found in nonfunctional regions of the genome, an assumption of a constant mutation rate is less likely to be violated by the occurrence of selection.

The phylogenetic analysis of SINE usually involves three basic steps:

1. identification/acquisition of PCR primers for SINE loci;
2. PCR amplification and electrophoretic visualization of size polymorphic bands corresponding to the presence (+) or absence (–) of fragments of target SINE inserts;
3. Southern hybridization (DNA-DNA blotting):
   a) blot of the PCR gel using a unit sequence of the SINE;
   b) blot of the PCR gel using the SINE flanking sequence.

The presence of SINE insertions (+ bands) define clades and the absence of these insertions for the same locus (– bands) is considered outgroups. A simple example of tree construction from PCR amplification analysis of SINE is illustrated in Figure 18.3.

18.5.2 Regulation of evolutionary change

The genetic and biochemical complexities are likely to have increased during evolution. Both DNA and proteins are dynamic molecules that undergo alternations and modifications. With two copies of a gene available in a genome, one copy could provide the necessary original function while the other accumulates mutations to alter its function. If this altered copy evolved eventually to serve a new function, it would tend to be retained in the genome and passed on to later generations. Many genes/proteins are homologous and are probably the products of gene duplication comprising a gene family. Some of factors inducing/regulating evolutionary changes are:

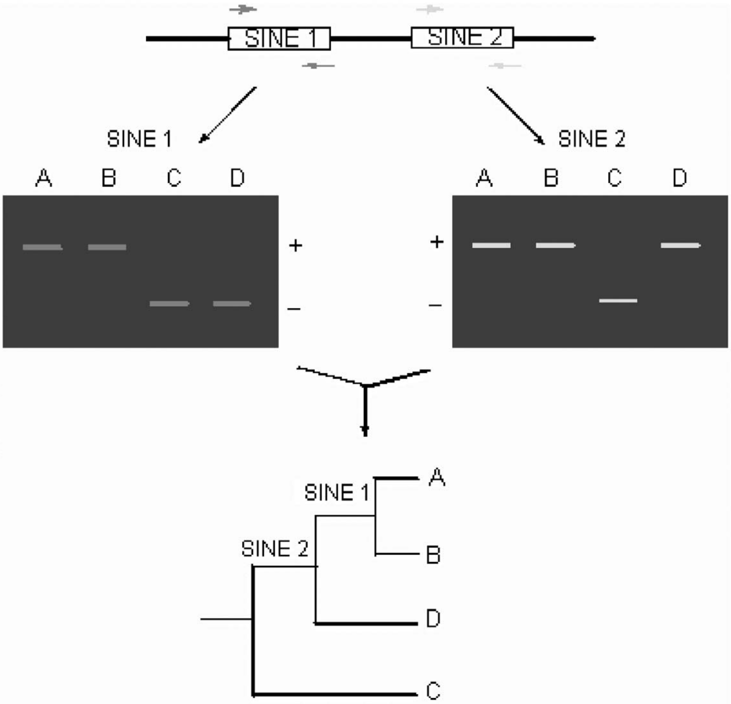


Figure 18.3 PCR Amplification analysis of SINEs and cladogram construction. The cladogram construction by PCR amplification analysis of two different SINE loci (SINE1 and SINE 2) is exemplified. Forward and reverse PCR primers (arrows) are designed to anneal at sites flanking a SINE inserted into the host genomes at a specific locus. Fragments for a given SINE are amplified by PCR for the different host taxa (A, B, C and D) being examined. Electrophoretic analysis of PCR products detects the presence (+ expected fragment size) or absence (– other). The cladogram that corresponds to the electrophoretic patterns is constructed. Taken from Shedlock and Okada (2000)

18.5.2.1 Provision of a selective advantage offers the ability to regulate change (McClintock, 1984). Genomes can evolve a heritable response to recurrent classes of environmental challenges in an increased probability. However, a good genomic strategy is to maintain a rate of change that is relatively low when an organism is well adapted to the stable environment, while being able to regulate the extent and type of genetic change in response to stress. Indeed a change in an organism's environment may induce a shifting array of metabolic enzymes, and alter the balance between genetic stability, repair and exploration.

18.5.2.2 Regulation of genetic change by methylation (Rossignol and Colot, 1999). DNA methylation via increasing/decreasing the extent of genome methylation or by altering the activity of DNA methylation enzymes, is a potentially global mechanism for genetic variation including recombination. A marked methylation allows regulation of the balance between faithful, random and mutagenic repair. It may also control the movement of transposons.

18.5.2.3 Transposable elements as molecular evolutionary force (Fedoroff, 1999). It has been estimated that a third of the human genome consists of repetitive sequences, mostly transposons and retrotransposons. Selective pressure appears to cause repetitive sequences and cleavage sites for mobile elements to be located in introns and untranslated regions. In addition, several biochemical mechanisms such as proteins binding to specific DNA sites, can connect mobile element insertion with gene function. For example, integrons can spread pathogenicity islands and antibiotic resistance in prokaryotic genomes and repetitive elements derived from RNA appear to have played a major role in eukaryotic evolution.

18.5.2.4 Role of RNA intermediates (Herbert and Rich, 1999). The mutation rate of a genome is likely to increase when genetic information is passed through RNA whether RNA is a viral genome or a retrotransposon because RNA polymerase reaction is neither edited nor subject to post-replicative repair. In addition, hotspots of a genetic chain in RNA retrotransposons can result from nonrandom patterns of a decreased fidelity strand transfer to other templates and untemplated extensions.

18.5.2.5 Induction of double-strand breaks (Holbeck and Strathern, 1999). Genetic variation can be increased by an induction of double-stranded breaks in the DNA. The recombination/repair of DNA helix breaks can increase the frequency of nearby base substitutions and frame shifts. For example, V-region hypermutation in the antibody appears to involve the focused generation of DNA breaks at a specific stage in B-cell development.

18.5.2.6 Horizontal transfer of genetic information (Kidwell, 1993). Bacteria can directly sample and adapt information available in their environment by taking up and incorporating DNA. Integrated genes move among bacterial genomes, carrying antivirotic resistance, pathogenicity and other properties. When favorable genetic changes arise on episomes that have been exposed to high mutation rate, these changes are poised to rapidly spread through a population. The horizontal transfer of gene clusters between organisms appears to be common and may also occur across species barriers.

18.5.2.7 Germ-line mechanisms of variation (Schwacha and Kleckner, 1997). During meiosis, mechanisms are induced to increase recombination that appear to bias the result or recombination towards the exchange of information rather than resolution to the unaltered DNA strands.

18.6 EVOLUTION OF PROTEIN STRUCTURE AND FUNCTION

The common ancestry or homology of proteins is usually inferred from similarity in sequences and/or structures. The determination of the evolutionary relationship by sequence similarity is straightforward for proteins of high sequence identity in the daylight zone (above 50% identity). Pair-wise sequence comparisons have difficulty in recognizing their relatives in the twilight zone (20–30% identity) where sequence profiles using a set of known homologues from a protein family are needed. In the midnight zone (below 20% identity), sequence comparison becomes unreliable and often structure data are necessary to recognize relatives. Substantial similarities in 3D structure may exist in the absence of significant sequence identity. Conversely, proteins may become dissimilar by evolutionary processes that their common origin cannot be detected from their sequences, even though they may still fulfill basically the same function. However, protein structures diverge much slower to provide evidence of common ancestry long after their sequence similarity has decayed.

Protein structure is more conserved than sequence. During evolution, the sequence can change considerably down to about 20–30% identity (twilight zone) before any large structural change occurs. Below about 20% sequence identity (midnight zone), however, a much more dramatic change is normally observed, which may correspond to mutations in critical residues for the function and/or structural stability of proteins. The extent of structural change for a certain amount of sequence change known as structural plasticity (or mutation sensitivity) varies with protein families. Plasticity does not appear to be correlated with structural class or mutation rate for the family. The structural diversity does not seem to show clear trends with functional diversity, though removal of functional constraints may allow more structural change provided the stability of the protein is maintained. In addition, physicochemical constraints imposed by the fold topology may also restrict structural diversity. For example, the mainly β-sandwich-like architectures appear more amenable to structural change than α/β barrel-like architectures (Orengo *et al.*, 2001).

18.6.1 Evolution of protein complexity: General

Protein structures evolve through a combination of mechanisms that often include gene duplication, followed by mutation and selection. A major genetic event in the evolution of protein structures is duplication. Gene duplication is central to the diversification of proteins. It is an effective path to increased complexity for both protein structure and function. Partial gene duplication has also contributed to function diversification by generating multiple tandem repeats or the occurrence of homologous domains in different domain architectures. Domain duplication and acquisition then lead to variations in the tandem arrangement of domains along a sequence. Duplication accompanied by gene fusion, is essential for a variety of other processes that result in the generation of novel proteins, such as unequal recombination (Marcotte *et al.*, 1998), circular permutation (Lindqvist and Schneider, 1997) and domain shuffling (Doolittle, 1995). Unequal recombination is the primary mechanism that gives rise to repetitive proteins. Circular permutation is the process by which N- and C-terminal deletions in a duplicated protein can result in a structure that appears to have its C-terminal part permuted to the N-terminus. Domain shuffling is the main mechanism for the rapid generation of novel domains by combining a set of modular domains. An important effect of domain shuffling is that proteins that are not homologous globally may contain homologous domains.

Protein domains are compact polypeptide structures generally organized around a clearly recognizable core and associated with a specific function or activity. Domains sharing the same fold display essentially the same 3D folding pattern and topology of core secondary structure elements. Since domains adopt only a limited number of folds (Salem *et al.*, 1999), the general structural similarity can be described in terms of protein folds. Proteins are considered to possess the same fold if they have the same major secondary structural elements in the same topology (orientation and connectivity). It is noted that proteins of the same fold do not necessarily share a common ancestor. Major structural similarity could arise independently due to a limited number of acceptable spatial arrangements of secondary structure elements (Pittsyn and Finkelstein, 1981). Furthermore, there exist evolutionarily related proteins that contain major structural differences and thus these proteins could be attributed to different folds.

Within the recognized folds, the proteins could be further grouped into one or more superfamilies, which are monophyletic assemblages characterized by a sequence signature and/or structural features unique to the constituent members. In turn, superfamilies are usually divided into families, compact sets of homologous proteins that share significant sequence similarity. Superfamilies are made of homologous families, i.e. they represent the result of divergent evolution, whereas folds typically group together analogous, i.e. convergently evolved superfamilies.

The significant sequence similarity is almost always reflected in local structural resemblance in the regions of conserved sequence motifs, which would hint at an evolutionary relationship. Weak sequence and structural similarities can be strengthened by a functional connection. Because similar folds can arise independently in evolution, structural similarity alone does not provide sufficient evidence of common ancestry. Additional factors such as retention of unusual structures, common domain organization and functional similarity must be taken into consideration to evaluate evolutionary relatedness. Due to the evolutionary tendency to conserve domain architecture, the co-occurrence of domains increases the probability of homology between proteins. Typically, multidomain proteins have a major domain that is the principal functional unit. This domain tends to be more conserved and shares significant sequence similarity among homologs.

Domains with identical or similar biochemical activities do not necessarily have the same fold. Conversely domains with different biochemical activities may group within the same fold. Recognition of some distant evolutionary events (such as unequal recombination, circular permutation and domain shuffling) has gradually made it possible to describe the basic complement of protein domains that was present in the last common ancestor (Aravind *et al.*, 2002).

18.6.2 Evolution of protein complexity: Domain duplication

The importance of domain duplication in evolution has long been recognized. In prokaryotes, at least 70% of the domains have been duplicated, whereas the figure is as high as 90% in eukaryotes (Chothia *et al.*, 2003). Since a large proportion of genes (up to 90% in eukaryotes) compose multidomain proteins while the diversity of proteins that can be assembled by duplicating domains and then combining them in different way is quite astonishing. Thus the domain can be viewed as a fundamental unit of evolution.

Analysis of completed genomes indicates that about two-thirds of the sequences can be assigned to as few as 1400 domain families for which structures are known. About 200 of these domain families are common to all kingdoms of life and account for nearly 50% of domain structure annotation in the genomes (Orengo and Thornton, 2005). During the course of evolution, proteins derived from a common ancestral protein can change their sequences via mutations, substitutions and/or InDels giving rise to families of homologous proteins. The protein structure data have provided insights into the mechanism by which domain duplication followed by divergence and/or domain fusion events have modulated the functions of the proteins. Thus domain duplication and modifications of their functions can expand the functional repertoire of the organism.

Most of the highly recurring domains are performing important generic functions. The extensive recurrence of domains is most likely due to the reuse of functionally important domain modules. Furthermore, highly duplicated domains have combined in different ways with many different domain partners, giving rise to the wide variety of multidomain proteins, which tend to have different functions. Alternation in domain partnership is a frequent mechanism for inducing functional change usually by altering the nature and the geometry of the active site of enzymes that affect substrate specificity while conserving (or semi-conserving) the reaction specificity (Todd *et al.*, 2001). The varied multidomain composition must be partly responsible for the diverse functions and phenotypes observed in different organisms.

18.6.3 Evolution of protein structure: Fold change

The structural evolution by gene duplication and fusion results in many of the protein structural domains containing symmetrical superfolds with pseudo-twofold symmetry (Ponting and Russell, 2002). For example, the $(\beta/\alpha)_8$ barrels of histidine biosynthetic enzyme (HisF) display eightfold pseudo-symmetry suggesting structural symmetry between the N-terminal (HisF-N) and C-terminal (HisF-C) halves of the $(\beta/\alpha)_8$ barrel scaffold. The two halves, HisF-N and HisF-C share sequence similarity and superimpose with a root-mean-square deviation (RMSD) of 1.58 Å (Lang *et al.*, 2000). Co-expression *in vivo* and joint refolding *in vitro* of HisF-N and HisF-C half-barrels produce a catalytically active enzyme (Hocker *et al.*, 2001). The sequence similarity between HisF-N and HisF-C and their ability to oligomerization support the notion that each half-subunit most likely descended from a common ancestor of homodimeric half-barrels.

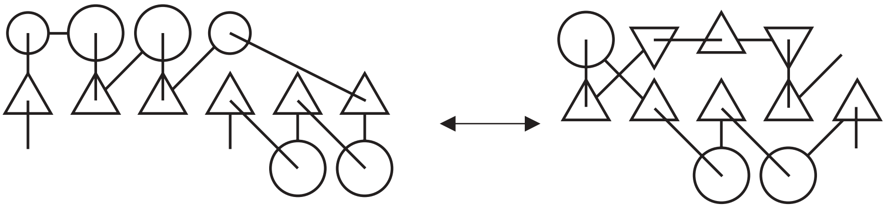
The protein structural similarity is generally described in terms of folds. However, evolutionarily related proteins may contain different folds, and proteins sharing the same fold do not necessarily descend from a common ancestor. Structural similarity may occur due to the limited number of favorable arrangements for secondary structure elements in fold space. Four of the most common mechanisms for protein-fold change (Grishin, 2001; Lupas *et al.*, 2001) include:

18.6.3.1 Insertion/deletion and substitution of secondary structural elements (Figure 18.4A). Insertions/deletions and substitutions are the most common events in protein evolution. Together they may lead to potentially significant changes in protein structures. Typically InDels are short (one to five residues), although longer InDels that occur at the periphery of the structure do not apparently affect its fold. However, consecutive InDels may cause substitutions of secondary structural elements (α-helix ↔ β-strand) and therefore protein folds. For example, lactate dehydrogenase (LDH) and NADH peroxidase (NHPOx) shares the common FAD/NAD binding motif (GxGxxG). The conversion of a connector helix C in LDH with 3 stranded β-meander (^{215}ERYEGDGRVQKVVTDKNAY^{233} in NHPOx) in NHPOx results in different ββα architecture (versus αβα) and thus different folds for these homologs.

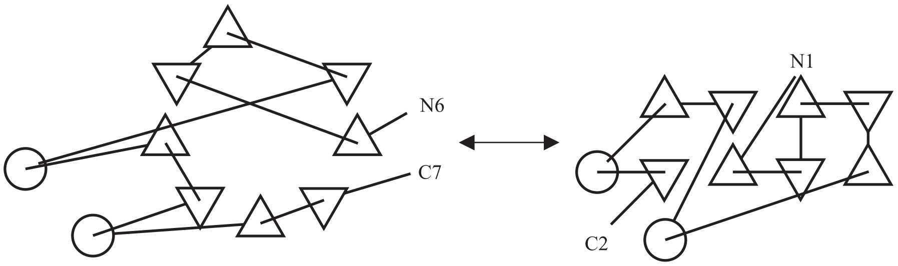
18.6.3.2 Circular permutation (Figure 18.4B). Amino and carboxyl termini of protein domain structures are frequently placed in close proximity, which favors circular permutations via ligation of the termini and cleavage at another site (Lindqvist and Schneider, 1997). The circular permutations do not change the spatial arrangement of secondary structure elements, but alter connectivity between secondary structures and thus protein folds. For example, the C2 domain is a Ca^{2+} binding module for proteins of signal transduction. The topologies of the C2 domain of phospholipase C_δ and synaptotagmin I are related by a circular permutation of their single β-strand, which covers about 12% of the protein chain.

18.6.3.3 β-Strand invasion/withdrawal (Figure 18.4C). The insertion of a β-strand(s) into existing β-sheets that requires hydrogen bond breakage at the withdrawing site and formation of the hydrogen bonds at the insertion site is termed strand invasion. This rearrangement of hydrogen bonding network in the β-sheets may drastically alter the protein structure if occurred in the conserved core regions. For example, all P-loop nucleotide triphosphatases (NTPases) share αβα sandwich architecture with the central mainly parallel β-sheet composed of βα units. However, connectivity between these βα units is different among NTPases families. The difference among the arrangements of the

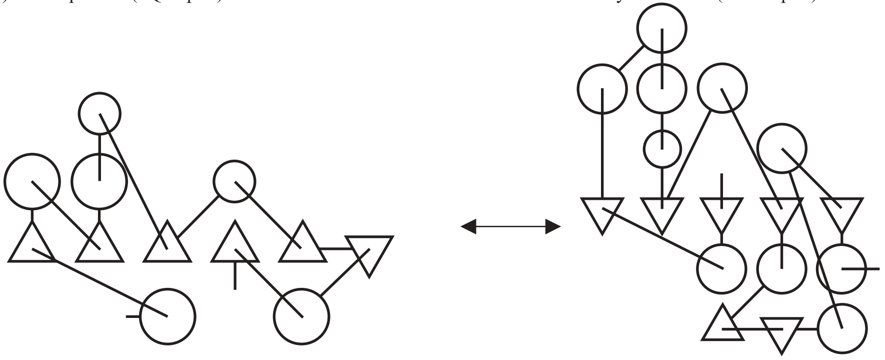
(a) Lactate dehydrogenase (1LDN.pdb), domain 1 NADH peroxidase (1NPX.pdb), domain 4



(b) Phospholipase C_δ (1QAS.pdb), B3 Synaptotagmin I (1RSY.pdb), C2



(c) Ras protein (1Q21.pdb) Adenylate kinase (2AKY.pdb)



(d) Retinol-binding protein (1HBQ.pdb) Thrombin inhibitor triabin (1AVG.pdb)

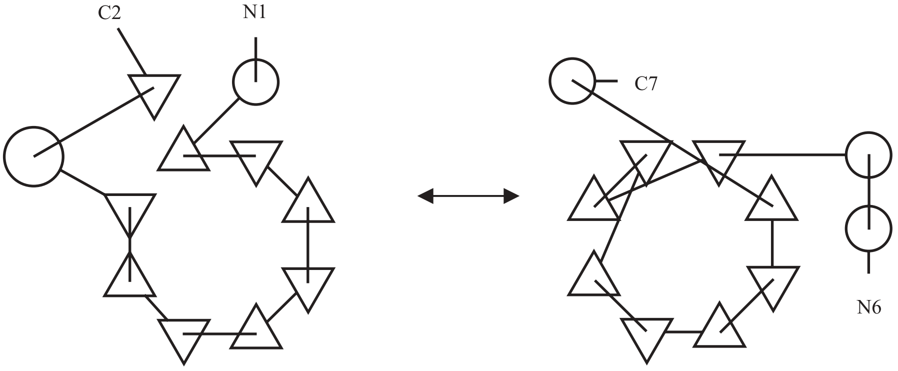


Figure 18.4 Common mechanisms for protein fold changes. TOPS representations for four of the most common mechanisms for protein fold changes involving secondary structures are illustrated with examples. They are (A) insertion/deletion/substitution between lactate dehydrogenase and NADH peroxidase, (B) circular permutation between phospholipase C_δ and synaptotagmin, (C) β-strand invasion/withdrawal between ras protein and adenylate kinase, and (D) β-hairpin flip/swap between retinol-binding protein and thrombin inhibitor triabin

β-sheet regions of adenylate kinase and ras protein can be rationalized in terms of strand invasion/withdrawal.

18.6.3.4 β-Hairpin flip/swap (Figure 18.4D). Another structural rearrangement in β-sheets involves internal swapping of the β-strands, which are commonly observed for the two adjacent β-strands forming a β-hairpin. Such a rearrangement creates (or removes) crossing loops and may generate unusual β-sheet topologies. For example, retinol-binding protein and thrombin inhibitor triabin shares significant sequence similarity with a domain of an 8-stranded up-and down β-barrel in which one β-hairpin flips around its axis (or swaps two β-strands with each other).

It seems likely that evolutionarily conserved domains that contain few InDels are homologous. Evolutionarily plastic proteins, in which only a few structural segments are functionally crucial, may sustain considerable evolutionary drift. For example, most of the structural segments in lysozyme-like proteins differ between family representatives and do not necessarily need to be homologous. Homologous proteins can have different folds brought about by these fold changes.

18.6.4 Evolution of protein function: Catalytic site convergence versus divergence

Structures and functions of proteins change in the succession of generations through random drift and natural selection. The basic processes of mutation, duplication and shuffling have led to a set of ancestral domains of proteins observed today. Domains associated with specific, critical functions tend to evolve under strong purifying selection forces and consequently show little diversity in sequence or domain architectures. Those domains that adopt catalytic functions typically show sequence conservation in the vicinity of the active sites but diverge in other regions of the molecule, producing new substrate specificities. In contrast, domains associated with complex regulatory functions tend to proliferate via multiple duplications, giving rise to a vast diversity of sequences and multidomain architectures. The emergence of complex regulatory functions during evolution seems to involve the recruitment of a versatile domain, followed by a positive feedback cycle, which allows the proliferation of the domain and results in the creation of a new related functional niche through the domain diversification.

Protein 3D structure is more conserved than either sequence or function. Different protein folds can converge on the same function, with identical catalytic residues built on completely different scaffolds. Conversely, a single architectural fold can adopt diverse sets of functions, with differing catalytic residues found in similar sites, similar catalytic residues found in different sites or the migration of catalytic residues within homologous protein folds. Features relating to function are the most obvious indicators of common ancestry. Evolutionary changes in functional/catalytic sites (Lupas *et al.*, 2001; Kinch and Grishin, 2002) can be considered with reference to:

18.6.4.1 Convergence of catalytic site in nonhomologous structures. Serine proteases represent the most notable example of evolutionary convergence in which one family is represented by the trypsin family, which includes chymotrypsin, trypsin and elastase and the other family by the bacterial protease, subtilisin. The two families of enzymes have similar function and mechanisms of action, even though they are not detectably related in structures. The active site of α-chymotrypsin consists of Ser-195, His-57 and Asp-102 forming the charge-transfer relay system or catalytic triad. Similar constellation of three such residues, Asp-32, His-64 and Ser-121, is also found in subtilisin where no sequence or fold homology to α-chymotrypsin is detectable (Kraut, 1977). Because of its catalytic importance, the same catalytic triad arose independently in subtilisin, presumably by convergent evolution.

Zinc-dependent proteases also built similar active sites into different structural templates during evolution. Two nonhomologous Zn-dependent proteases, thermolysin and mitochondrial processing peptidase (MPP) display remarkable structural and functional convergences (Makarova and Grishin, 1999). Thermolysin resembles a Rossmann fold and belongs to the family of zincin-fold proteases. This enzyme requires an essential Glu and a single zinc ion for its catalytic action. The catalytic Glu and two of three Zn ligands form a signature HExxH Zn-binding motif. By contrast, MPP possesses a modified ferredoxin-fold and the inverted signature pattern, HxxEH. Notwithstanding, all major functionally important structural elements are present in both thermolysin and MPP with different topological connections, e.g. the inversion of the signature motifs. Each structure displays a loose bundle of α-helices packed on one side of a five-stranded β-sheet with an overlap of secondary structural elements (superimposition with rmsd of 3.4 Å). Thus two different folds of thermolysin and MPP arose independently, converging on both similar overall structure and similar spatial albeit inverse orientation of catalytic/binding residues.

18.6.4.2 Divergence of catalytic site in homologous structures. The DNA polymerase superfamily, DNA-dependent RNA polymerase (DNA primase), RNA-dependent RNA polymerases (RDRP), reverse transcriptases (RT), and the nucleotide cyclases all share a common catalytic core, the palm domain. All these enzymes display similar metal-dependent catalytic mechanisms, suggesting that they evolved from a common ancestor. The core of the palm domain is a $(\beta\alpha\beta)_2$ unit, with two conserved acidic residues at the end of strand 1 and in the loop between strands 2 and 3 that chelate two divalent cations. Most polymerases have extensions to or insertions into the palm domain, the finger modules which ensure tighter interactions with the template polynucleotides. The palm domain of DNA and RNA polymerases appears to be a version of the RNA recognition motif (RRM)-like fold that is seen in ancient nucleic acid-binding domains. These observations suggest that the ancestral polymerase probably evolved from a nucleic acid-binding domain that might have functioned as an accessory to a self-replicating nucleic acid. The metal-binding active site probably evolved subsequently within the ancestral palm domain, along with the intrinsic catalytic activity. Finger domains appear to have been independently inserted, after the divergence of the major polymerase lineages, as adaptations for fine-tuning the polymerases for their specific template and biological functions (Aravind *et al.*, 2002). While the active site of eukaryotic primase (EP) occupy a similar spatial configuration, EP appears to have evolved differently from the classic palm domains with a highly distorted fourth strand of the $(\beta\alpha\beta)_2$ unit and an additional active acidic residue. Furthermore, EPs contain an acidic and basic residue pair in place of the two acidic residues of the palm domains of other polymerases.

A conserved catalytic triad, Cys-His-Asp/Glu (CHD/E) characterizes type I glutamine amidotransferase (GAT) domains. The GAT superfamily, which also includes carbamyl phosphate synthetase (CPS) and an intracellular protease from *Pyrococcus horikoshii* (PHP), contains a variation of the Rossmann fold topology with a central sheet formed by five parallel β-strands connected by flanking α-helices on either side. A large insertion of variable sequences is also common in GAT domains. The structure of a CPS small subunit contains a conserved catalytic CHE triad with His353 and Glu355 located within the loop connecting βE, αE and Cys269 in a sharp turn at the N-terminus of αD, referred as the nucleophile elbow. In PHP, the migration of catalytic residues occurs such that the catalytic Cys100 is retained in the same nucleophile elbow, with a neighboring His101. But Glu474 is provided by an adjacent subunit (Du *et al.*, 2000). Thus PHP has adapted distribution of its catalytic triad to different structural elements with reference to its GAT domain homologs and included oligomerization as a strategy for creating this diversity.

Generally, the primary sequence of a protein dictates its fold and function, and cumulative changes in this sequence lead to the evolution of protein structures and functions, although exceptions are known whereby similar sequence fold into different structures. Thus protein structures can evolve and change to generate new folds and topologies. Similarly, the functional evolution of proteins can be divergent, with a migration of catalytic residues within homologous folds, or it can be convergent, with identical functional residues forming similar spatial arrangements in a completely different protein folds.

18.7 REFERENCES

Aravind, L., Mazumder, R., Vasudevan, S. and Koonin, E.V. (2002) *Current Opinions in Structural Biology*, **12**, 392–9.

Brown, J.R. and Doolittle, W.F. (1997) *Microbiology Molecular Biological Reviews*, **61**, 456–502.

Caetano-Anollé, G. (2002) *Journal of Molecular Evolution*, **54**, 333–45.

Chothia, C., Gough, J., Vogel, C. and Teichmann, S.A. (2003) *Science*, **300**, 1701–3.

Czelusnlak, J., Goodman, M., Moncrief, N.D. and Kehoe, S.M. (1990) *Methods in Enzymology*, **183**, 601–15.

Dayhoff, M.O. (1978) *Atlas of Protein Sequence and Structure*, Vol. 5, suppl. 3. National Biochemical Research Foundation, Washington, DC.

Doolittle, R.F. (1995) *Annual Reviews in Biochemistry*, **64**, 287–314.

Doolittle, W.F. (1999) *Science*, **284**, 2124–8.

Dover, G.A. (1986) *Trends in Genetics*, **2**, 159–65.

Du, X., Choi, I.G., Kim, R. *et al.* (2000) *Proceedings of the National Academy Sciences, USA*, **97**, 14079–84.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998) *Biological Sequence Analysis*, Cambridge University Press, Cambridge.

Fedoroff, N.V. (1999) *Annals of New York Academy of Science*, **870**, 251–264.

Felsenstein, J. (1981) *Journal of Molecular Evolution*, **17**, 368–376.

Felsenstein, J. (1985) *Evolution*, **39**, 783–91.

Felsenstein, J. (1988) *Annual Review of Genetics*, **22**, 521–65.

Felsenstein, J. (1996) *Methods in Enzymology*, **266**, 418–27.

Fitch, W.M. (1979) *Syst. Zoology*, **28**, 375–9.

Fitch, W.M. and Margoliash, E. (1967) *Science*, **155**, 279–84.

Fitch, W.M., Bush, R.M., Bender, C.A. *et al.* (2000) *Journal of Heredity*, **91**, 183–5.

Garcia-Vallve, S., Romeu, A. and Palau, J. (2000) *Genome Research*, **10**, 1719–25.

Grishin, N.V. (2001) *Journal of Structural Biology*, **134**, 167–85.

Hendy, M.D. and Penny, D. (1982) *Math. Biosci.* **59**, 277–90.

Herbert, A. and Rich, A. (1999) *Nature, Genetics*, **21**, 265–9.

Higgins, D.G., Thompson, J.D. and Gibson, T.J. (1996) *Methods in Enzymology*, **266**, 383–402.

Higgs, P.G. and Attwood, T.K. (2005) *Bioinformatics and Molecular Evolution*, Blackwell, Malden, MA.

Hillis, D.M. (1987) *Annual Reviews in Ecological Systems*, **18**, 23–42.

Hillis, D.M. and Bull, J.J. (1993) *Systems in Biology*, **42**, 182–92.

Hillis, D.M., Allard, M.W. and Miyamotao, M.M. (1993) *Methods in Enzymology*, **224**, 456–87.

Hocker, B., Beismann-Driemeyer, S., Hettwer, S. *et al.* (2001) *Nature Structural Biology*, **8**, 32–6.

Holbeck, S.L. and Strathern, J.N. (1999) *Annals of New York Academy of Sciences*, **870**, 375–7.

Horn, F. Vriend, G. and Cohen, F.E. (2001) *Nucleic Acids Research*, **29**, 346–9.

Huelsenbeck, J.P. and Imennov, N.S. (2002). *Systems in Biology*, **51**, 155–65.

Kidwell, M.G. (1993) *Annual Reviews in Genetics*, **27**, 235–56.

Kimura, M. (1980) *Journal of Molecular Evolution*, **16**, 111–20.

Kinch, L.N. and Grishin, N.V. (2002) *Current Opinions in Structural Biology*, **12**, 400–8.

Kraut, J. (1977) *Annual Reviews in Biochemistry*, **46**, 331–58.

Lake, J.A. (1987) *Molecular Biology Evolution*, **4**, 167–91.

Lang, D., Thoma, R., Henn-Sax, M. *et al.* (2000) *Science*, **289**, 1546–50.

Lindqvist, Y. and Schneider, G. (1997) *Current Opinions in Structural Biology*, **7**, 422–7.

Lupas, A.N., Ponting, C.P. and Russell, R.B. (2001) *Journal of Structural Biology*, **134**, 191–203.

Makarova, K.S. and Grishin, N.V. (1999) *Protein Science*, **8**, 2537–40.

Marcotte, E.M., Pelligrini, M. and Yeates, T.O. (1998) *Journal of Molecular Biology*, **293**, 151–60.

McCalden, P. and Argos, P. (1988) *Proteins*, **4**, 99–122.

McClintock, B. (1984) *Science*, **226**, 792.

Miyamoto, M.M. and Cracraft, J. (eds) (1991) *Phylogenetic Analysis of DNA Sequence*, Oxford University Press, Oxford, UK.

Needleman, S.B. and Wunsch, C.D. (1970) *Journal of Molecular Biology*, **48**, 443–53.

Nei, M. and Kumar, S. (2000) *Molecular Evolution and Phylogenetics*, Oxford University Press, Oxford, UK.

Orengo, C.A., Sillitoe, I., Reeves, C. and Pearl, F.M.C. (2001) *Journal of Structural Biology* **134**, 145–65.

Orengo, C.A. and Thornton, J.M. (2005) *Annual Reviews in Biochemistry*, **74**, 867–900.

Page, R.D.M. (1996) *Computer Applied Bioscience*, **12**, 357–8.

Page, R.D.M. and Holmes, E.C. (1998) *Molecular Evolution: A Phylogenetic Approach*, Blackwell Science, Oxford, UK.

Paraskevis, D., Lemey, P., Salemi, M. *et al.* (2003) *Molecular Biology Evolution*, **20**, 1986–96.

Penny, D. and Hendy, M. (1986) *Molecular Biology Evolution*, **3**, 403–17.

Philippe, H. and Laurent, J. (1998) *Current Opinions in Genetic Development*, **8**, 616–23.

Phillips, A., Janies, D. and Wheeler, W. (2000) *Molecular Phylogenetic Evolution*, **16**, 317–30.

Pittsyn, O.B. and Finkelstein, A.V. (1981) *Quarterly Reviews in Biophysics*, **13**, 339–86.

Ponting, C.P. and Russell, R.B. (2002) *Annual Reviews in Biomolecular Structure*, **31**, 45–71.

Rossignol, J-L. and Colot, V. (1999) *BioEssays,* **21**, 402–11.

Saitou, N. (1996) Methods in Enzymology, **266**, 427–49.

Saitou, N. and Nei, M. (1987) *Molecular Biology Evolution*, **4**, 406–25.

Saitou, N. and Imanishi, T. (1989) *Molecular Biology Evolution*, **6**, 514–25.

Salem, G.M., Hutshinson, E.G., Orengo, C.A. and Thornton, J.M. (1999) *Journal of Molecular Biology*, **287**, 969–81.

Schwach, A. and Kleckner, N. (1997) *Cell*, **90**, 1123–35.

Shedlock, A.M. and Okada, N. (2000) *BioEssays*, **22**, 148–60.

Shedlock, A.M., Takahashi, K. and Okada, N. (2004) *Trends in Ecological Evolution*, **19**, 545–53.

Smith, T.F. and Waterman, M.S. (1981) *Journal of Molecular Biology*, **147**: 195–7.

Stagle, J.R. (1971) *Artificial Intelligence: The Heuristic Programming Approach*, McGraw-Hill, New York.

Todd, A.E., Orengo, C.A. and Thornton, J.M. (2001) *Journal of Molecular Biology*, **307**, 1113–43.

Wilson, A.C., Carlson, S.S. and White, T.J. (1977) *Annual Reviews in Biochemistry*, **46**, 573–639.

Woese, C. (1998) *Proceedings of the National Academy Sciences, USA*, **95**, 6854–9.

World Wide Webs cited

BCM:	http://dot.imgen.bcm.tme.edu..9331/multi-align/multi-align.html
BioEdit:	http://www.mbio.ncsu.edu/BioEdit/bioedit.html
BLAST:	www.ncbi.nim.nih.gov/entrez/query.fcgi?db=Taxonomy
CLUSTAL:	ftp://ft-igbmc.u-strasbg.fr/pub/
DDBJ:	http://www.ddbj.nig.ac.jp/
EBI:	http://www.ebi.ac.uk/
Institute of Genomic Research:	http://www.tigr.org/tdb/mdb/mbdcomplete.html
Joint Genome Research Institute:	http://www.jgi.doe.gov/JGI_microbioa/html/index.html
MALIGN:	ftp://ftp.amnh.org/pub/people/wheeler/malign/
NCBI:	http://ncbi.nlm.nih.gov/Genomes/index.html.
PHYLIP:	http://evolution.genetics.washington.edu/phylip.html
Phylogeny Programs:	http://evolution.genetics.washington.edu/phylip/software.html.
ReadSeq:	http://dot.imgen.bcm.tmc.edu:9331/seq-util/Options/readseq.html
Sanger Centre:	http://www.sanger.ac.uk/Projects/Microbes/
TREEALIGN:	ftp://ftp.ebi.ac.uk/pub/sofware/unix/treealign.tar.Z
TreeView:	http://taxanomy.zoology.gla.ac.uk/rod/treeview.html
WebPHYLIP:	http://sdmc.krdl.org.sg:8080/~lxzhang/phylip/welcome.html
Phylogenetic analysis software programs:	Table 18.4
Phylogeny databases and utilities:	Table 18.6

INDEX

Page references followed by t indicate material in tables.

A
AAindex database, 597
aa-tRNA, accommodation and rejection of, 481
Abbreviations, xvii–xxviii
  nucleic acid, 56
Ab initio approach, 250
Ab initio gene prediction, 572–575
Ab initio (*de novo*) protein structure prediction (AB), 618–619, 624–625
Absolute specificity, 328
Absorption, of radiation, 183–184
Absorption band, 187
Absorption spectroscopy, perturbation difference, 189–190
Abundance-based (analytical) microarrays, 638–639
Abzymes, 383–386, 511
Accelerated reactions, 323
Accessible surface, of folded structures, 130–133
Acetylation, 221, 484
Acid–base catalysis, versus metal ion catalysis, 392–393
Acid catalysis, 345–346
Acids, pectinic, 165. *See also* Amino acid entries
Activation energy, determination of, 343
Activation-substitution coupling reaction, 225
Activator ion/prosthetic group, introduction of, 505
Active site(s)
  attributes of, in enzymes, 330
  chemical modifications of, 349–350
Active site enzyme mechanism studies, 349–356
Activity-based probe (ABP), 640–643
Activity probe structures, 641
Acylation, 484
ADAM algorithm, 284
Adenylyl cyclase (AC), in signal transduction, 406–408
Adsorption, in biomacromolecule purification, 34
Adsorption chromatography, 36–37. *See also* Affinity chromatography
Advanced glycated end-product (AGE), 486
Affinity alkylating probes, 642
Affinity-based probe (AFBP), 640
Affinity chromatography, 38–40
  protein-ligand combinations for, 39
A-form DNA, 68, 72
  CD of, 213
A-form double helix (A-helix), 86
Agarose gels, 41, 42
Alanine, characteristics of, 19t
Alanine racemase, 371
Aldehyde microarray surfaces, 526, 527t
Algorithms, computer, 536
Alkylated nucleotides, dealkylation of, 459
Alkylating probes, 642
Alleles, 560
Alloisoleucine, 18
All-or-none transition, 269
Allosteric/cooperative kinetics, 340–341
Allosteric enzymes, inhibitor/activator effect on, 380
Allosteric model, of homotropic interactions, 296–297
Allosteric regulation, structural basis of, 381–383
Allosterism, 378–380
  in biomacromolecular interaction, 295–299
Allothreonine, 18
αα hairpin, 117, 308–309
α-amino acids. *See also* Amino acids
  characteristics of, 19–21t
  ionization behavior of, 21–22
  as protein constituents, 18–23
  pyridox phosphate mediated chemical transformations of, 370t
α-β domain structures, 121
α + β proteins, 124, 135
α/β proteins, 135
α-chymotrypsin, crystal structure of, 368
α domain structures, 120
α-helical conformation, in polypeptide chains, 273
α-helical folds, 122–123
α-helical proteins, 133
α-helical structure, 111–113
α-helix (helices)
  CD curve of, 211
  Cotton effects for, 210–211
  in globular proteins, 116–117
  ORD spectrum of, 212

Biomacromolecules, by C. Stan Tsai
Copyright © 2007 John Wiley & Sons, Inc.

  properties of, 112–113
  role in recognition, 306–307
α-hydroxy acids, 646
Alternative splicing, 566
AMBER package, 264
  molecular mechanics energy minimization using, 265–267t
American Type Culture Collection (ATCC), 496
Amide I/II bands, 195–196
Amide binding site positions, 319
Amide chromophore, 212
Amine-imine equilibrium, 16
Amine microarray surfaces, 526, 527t
Amino acid analogs, 644, 645–646t
Amino acid composition, analysis of, 96–97
Amino acid functional groups, chemical modifications of, 351–352t
Amino acid pairs, distinguishing, 106
Amino acid replacement, 681–682
Amino acid residues, 187–188
  chemical modification sites in, 350
  encoded, 684
  hydrophobicities and related parameters of, 280t
  identification of, 633–634
Amino acid residue side chains, properties of, 18–21
Amino acids. *See also* α-amino acids
  adenylylation of, 372
  common, 105t
  as GPI attachment sites, 169
  hydrogen bonding of, 315
  incorporation into protein, 647
  spectral parameters for, 188t
Amino acid sequences, 598
  searching, 601
  steps in determining, 96
  structure prediction from, 276–282
Aminoacylation, specificity of, 372–374
Aminoacyl-tRNA, synthesis of, 472–473
Aminoacyl-tRNA binding, 477
Aminoacyl-tRNA synthetases (aRS's), 372, 472
  aminoacylation mechanisms catalyzed by, 373
  binding specificity of, 311
  characteristics of, 474t
Aminodicarboxylic acids, in peptide synthesis, 232–233
Aminopeptidases, 427
Amino protecting groups, 222–223t
Ampholytes, 21
Amylopectin, 164
Amylose, 162, 163–164
Anabolism, 399
Anchoring groups, on polymeric supports, 230
Ancient conserved regions (ACRs), 576
Angiogenesis, blocking, 510
Angle strain, 11
Animal lectins, 313–315
Anion exchangers, 37
Anomeric configuration, 238
Antibodies
  antiphosphonamidate and antiphosphonate, 383–384
  diversity of, 506
  as entropy traps, 384
  flexibility of, 303
  selection by phage display, 507
  specificity and affinity in, 508
  structure of, 300–303
  therapeutic, 508–511
Antibody-antigen complexes, 303–305
  changes in surface areas and contacts in, 304t
Antibody-antigen interactions, 300–305
  specificities of, 304–305
Antibody binding energy, 384–385
Antibody catalysis, selective, 383
Antibody-combining site, genetic or chemical modification of, 386
Antibody-directed enzyme prodrug therapy (ADEPT) technology, 510
Antibody efficacy, enhancing, 510
Antibody engineering, 506–511
Antibody-forming cells, 302
Antibody response, 300
Anticodons, 473
Anticodon stem, in tRNA, 83
Antiparallel β-sheets, 114
Antiphosphonamidate antibodies, 383–384
Antiphosphonate antibodies, 383–384
Antisense RNA, 85, 500
AOBase Web site, 500
A-platforms, 88
Apoptosis, 419–422
  pathways in, 421
  regulation of, 422
Apoptotic protease activating factor-1 (Apaf-1), 421
Aptamers, DNA/RNA as, 90–91
Arabidopsis thaliana database, 605
Arginase, 368
Arginase catalyzed hydrolysis, 369
Arginine, characteristics of, 20t
Arithmetic Logic Unit (ALU), 534
A-RNA, 81
Aromatic side chains, 303
  absorption and, 188
Asialoglycoprotein receptor (ASGPR), 318
Asparagine, 303
  characteristics of, 19t
Aspartate aminotransferase, 370
Aspartic acid, characteristics of, 20t
Assays, microarray, 530
Asymmetric PCR, 498. *See also* Polymerase chain reaction (PCR)
Atomic coordinate files, reliability characteristics of, 218t
Atomic coordinates, for three-dimensional structures, 218
ATP-dependent protein degradation, 430–433
ATP hydrolysis, 432

A · T(U) base pairing, 275
AUGUSTUS, 575
AutoDock tool, 283, 284
Automated amino acid analyzer, 96–97
Automated DNA sequencing technology, 62–63
Automatic docking, 283
Axial rise, 65

B
Bacterial genome, 559
Bacteriophages, DNA in, 443
Ball-and-stick models, computer, 537–540
Base catalysis, 345–346
Base excision repair (BER), 459–460
Base excision repair pathway, 460
Base-pairing patterns/schemes, 90
  as models for approximate secondary structures, 281–282
Base pairings
  locally disrupted, 73–74
  in nucleic acids, 275
Base pair roll, 65
Base-pair structure, recognition of, 480–481
Base pair tilt, 65
Bases, availability for protein recognition, 307
Base stacking, 89–90
Basic Local Alignment Search Tool (BLAST) program, 521–523, 599. *See also* BLAST searches
B cells, 302
Bcl-2 homology (BH) regions, 422
βαβ motif, 118, 123
β-α-β proteins, 135
β barrel motif, 118
ββ hairpin, 118
β-bulge, 114
β domain structures, 120–121
β-elimination reaction, 53
β-hairpin flip/swap, 706
β hairpins, 115
β-pleated structures, 112
β-sheet conformations, CD spectra of, 212
β sheet folds, 123
β sheet packings, 123
β-sheet proteins, 133–135
β-sheets, 211, 301
  in globular proteins, 117
  structure of, 113–115
β-strand invasion/withdrawal, 704–706
B-form DNA, 66–68, 70–73
  CD of, 213
  sequence dependent modification to, 71–73
Biantennary glycans, 171
Bi bi reaction, steady-state treatment of, 339
Bidirectional DNA replication, 447
Bimolecular reactions, susceptibility to proximity effects, 384
Binding
  multiple-site, 292
  single-site, 291
Binding equation, derivation of, 297
Binding parameters, experimental evaluation of, 293t
Binding plots, responses of multiple-site receptors to, 295t
Binding proteins, 453–454
Binding sites, introduction of catalytic groups or cofactors at, 385–386
Biocatalysis/Biodegradation Database (UM-BBD), 399
Biocatalysts
  assessing effectiveness of, 322–323
  definition and classification of, 322–325
Biochemical compounds, three-dimensional models of, 28
Biochemical databases, 549–551
Biochemical equilibria, 289
Biochemical oscillation, 377
Biochemical polypeptide chain ligation, 245
Biochemical processes, noncovalent forces in, 5
Biochemical spectroscopy, 183–185
Biochemical synthesis, 220–248
Biochemistry/molecular biology
  applications of NMR in, 197–198
  Internet resources for, 546–547
Biochips, 525. *See also* Microarrays
Biocomputing freeware, 537t
BioEdit Web site, 689
Bioengineering, of biomacromolecules, 494–511
Bioinformatics. *See also* Biomacromolecular informatics; Informatics; Proteome informatics
  protein structure analysis using, 616–629
  sites concerning, 13
Biological activity, hydrophobic facilitation of, 360
Biological sciences, Internet as a resource for, 546–548
Biomacromolecular binding, predicting, 282–285
Biomacromolecular catalysis, 322–397
  enzymes and, 325–333
Biomacromolecular characterization, 44–53
  spectroscopic methods used in, 186t
Biomacromolecular crystals, preparation of, 216
Biomacromolecular evolution, 680–709
  molecular phylogeny, 685–687
  phylogenetic analysis of biosequences, 687–693
  variation in biomacromolecular sequences, 680–685
Biomacromolecular informatics, 515–557. *See also* Bioinformatics; Biosequences; Informatics; Proteome informatics
  computer technology and, 533–548
  databases, 548–553
  gene ontology, 553–554

Biomacromolecular interactions, 289–321
  allosterism and cooperativity in, 295–299
  antibody-antigen, 300–305
  binding site and, 290
  multiple equilibria in, 291–295
  nucleic acid, 305–311
Biomacromolecular properties
  predicting, 249
  representing, 264
Biomacromolecular sequences, variation in, 680–685. *See also* Biosequences
Biomacromolecular structure(s), 11–12, 55–93. *See also* Biomolecular structure entries; Polysaccharides; Protein structures
  abstracting from sequences, 276
  chemical synthesis of, 220–248
  classification and structures of RNA, 81–85
  conformational energetics, 143–144
  conformational maps, 108–110
  DNA secondary structure and structure polymorphism, 63–77
  nucleic acid applications, 90–91
  nucleic acid structure energetics, 89–90
  organizational levels of, 177–180
  RNA folds and structure motifs, 86–89
  sequence analysis of nucleic acids, 57–63
  supercoiling and tertiary structure of DNA, 77–80
  3D information for, 217
  3D modeling of, 249–250
Biomacromolecular systems, forms of entropy in, 262
Biomacromolecule chromophores, spectroscopic observation of perturbations of, 189–190
Biomacromolecule-ligand interactions, 296
Biomacromolecule purification, 31–44, 45t
  bulk, 34
  chromatography in, 34–40
  electrophoresis in, 40–44
  methods of, 33–34
  procedures in, 35t
  steps in, 31
Biomacromolecules, 1. *See also* Biomacromolecular entries; Biomacromolecule purification; Biomolecules; Biopolymers; Biosequences
  absorption spectroscopy of, 187
  bioengineering of, 494–511
  buoyant density of, 52
  charge groups in, 290
  classes of, 2t
  comparison of, 12t
  configuration of, 9–10
  conformations of, 10–11
  cooperative structural transition of, 270
  covalent bonds in, 4–5
  covalent linkages of monomer units in, 56
  crystallographic study of, 216–219
  detection of changes in, 291
  dimensions of, 50–51, 52t
  final coordinates of, 216
  folding of, 262, 491–494
  interest in, xiii
  MolD simulation of, 259–260
  molecular weight of, 44–50
  monomer constituents of, 16–30
  optical activity of, 212
  sensitivity to chemical reactions, 221
  in solution, 289–291
  treatment of chain conformation problems in, 268
BioMagResBank (BMRB), 205
Biomolecular interaction network database (BIND), 291
Biomolecular Structure and Modeling Group Web site, 326
Biomolecular structure determinations, via NMR, 202
Biomolecular synthesis, direct condensation for, 229t
Biomolecules
  electronic behavior in, 254
  electrospray ionization of, 50
Biopolymers
  biosynthesis of, 436
  solid phase synthesis of, 226
Biorhythmicity, 377
Biosequences, 515–525
  evolution of, 697–701
  phylogenetic analysis of, 687–693
Biosynthesis
  of glycoproteins, 437–442
  of oligo- and polysaccharide chains, 436–437
  saccharide, 436–442
Bioverse Web site, 613
Bispecific antibodies, 510
Bit score, 524
BLAST searches, 619. *See also* Basic Local Alignment Search Tool (BLAST) program
BLOCKS database, 603–604, 615
B lymphocyte receptor, 419
Boc/Bzl protecting groups scheme, 234–235
BODIPY series, 530–531
Boehringer Mannheim metabolic pathway chart, 398
Bond angles, 4
Bonding interactions, 251
Bond length, 4
Bottom-up clone library, 564–565
Boutique databases, 598, 605, 606t
Bragg equation, 215
Branched glycans, 13
  linear codes for, 152–153
Branched oligosaccharides, 149
Branched polysaccharides, 163
BRENDA Web site, 324
Browsers, 544, 545
Buckingham potential, 6, 8
Buffer electrofocusing, 43
Bulged bases, 74

Bulged-G RNA motifs, 87
Bulge-helix-bulge RNA motif, 88
Bulk biomacromolecule purification, 34
Buoyant density, of biomacromolecules, 52

C
C2 axis, 69
C2 domains, 411
C_2—H_2 zinc finger motif, 309
C4′—C5′ orientation, 18
Ca^{2+}/calmodulin-dependent kinase (CaM kinase), 413. *See also* Calcium entries
Ca^{2+}-dissociation constants, 411
Ca^{2+}-induced Ca^{2+}-release (CICR), 410
Ca^{2+}-sensors, 402
CADB Web site, 108
Calcium binding proteins, 411
  intracellular, 412t
Calcium signaling, 410–414. *See also* Ca^{2+} entries
  nitric oxide in, 413–414
Calmodulin, 411–413
Cambridge Crystallographic Data Centre (CCDC) database, 252
CaM kinase II, 413
cAMP-response element binding protein (CREB), 415. *See also* 3′,5′-Cyclic adenosine monophosphate (cAMP)
Cancer, monoclonal antibodies against, 510. *See also* Carcinogenesis
Cancer databases, 590t
Canonical base pairings, 63
Capillary electrophoresis, 42
CarbBank database, 173, 662
Carbohydrate analyses
  chemical methods for, 154–155
  enzymatic methods for, 155–157
  spectrometric methods for, 157–161
Carbohydrate-binding proteins, major groups of, 315
Carbohydrate-lectin interaction, molecular recognition in, 312–320
Carbohydrate metabolism, 436
Carbohydrate microarrays, 529
Carbohydrate recognition, pnmary monosaccharide specificity in, 318–319
Carbohydrate-recognition domains (CRDs), 314, 315
Carbohydrate-recognizing lectins, 314–315
Carbohydrates
  Cotton effects of, 213
  IR spectra of, 197
Carbohydrate Structure Suite (CSS), 662
Carboxyl protecting groups, 226t
Carboxypeptidase A, 427, 429
Carcinogenesis, telomerase re-activation and, 455. *See also* Cancer entries
Cascade effect, 375–377
Caspase-activated DNase (CAD), 422
Caspase-mediated apoptosis, 421
Caspases, 420–421
CaspR Web site, 627
Catabolism, 399
Catalysis, by retaining glycosidases, 424
Catalytic antibodies, 325, 511
Catalytic efficiency, of enzymes, 325, 328
Catalytic enhancement, by multienzyme complexes, 332
Catalytic groups, introduction at binding sites, 385–386
Catalytic proteins. *See* Enzyme entries
Catalytic reaction equilibrium, shift in, 505
Catalytic RNA, characteristics of, 386–388
Catalytic site/active site information, online, 330
Catalytic Site Atlas (CSA), 615–616
Catalytic sites
  convergence of, 706–707
  divergence of, 707–708
Catenanes, 80
CATH protein domain database, 136, 138t. *See also* Class, Architecture, Topology, Homology (CATH) database
Cation exchangers, 37
CAZY Web site, 666
cccDNA, 78
C-DNA, 69
cDNA libraries, 565
  sequencing, 583
CD spectra, 209–212. *See also* Circular dichroism spectroscopy
Cell disruption methods, 31, 32t
Cell lysis, 31, 32t
Cell-map proteomics, 594
Cell surface proteins, 136
Cellular proteins, degradation of, 430
Cellulose, 162, 165–166
Cellulose chains, tertiary structure of, 166
Center for Biological Sequence Analysis (CBS), 616
Central nervous system (CNS) receptors, 403
Central processor unit (CPU), 533, 534–535
Centre de Recherches sur les Macromolécules Végétales (CERMAV), 663
Centrifugation, 31–32
  in biomacromolecule purification, 34
  CsCl density gradient, 52
Cereal lectin, 312, 315
Chain elongation, 463, 477–479
  in peptide synthesis, 235
Chain initiation, in protein translation, 475–477
Chain termination, 463, 479
Chain terminators, 60
Chair conformations, 24
Character-based phylogenetic method, 691–692
Chargaff’s rules, 66
Charged biomacromolecules, 41
Charge groups, in biomacromolecules, 290
“Charge relays,” 327
Chem3D package, 264
ChemFinder Web site, 28

Chemical bonds, molecular mechanical treatment of, 254–255
Chemical bond stretching, potential energy profile for, 254–255
Chemical cleavage DNA sequencing, 58, 59–60
Chemical derivatization, 158–159
Chemical enzyme reaction mechanism studies, 344
Chemical ligation, 245
  in peptide synthesis, 235–236
Chemical modifications, active site, 349–350
Chemical precedent approach, 357
Chemical shifts, 199–200
  for amino acids in random coil, 205t
Chemical synthesis, of biomacromolecular structures, 220–248
Chemical transduction, 398–400
Chirality, 9
Chitin, 166
Chitosan, 166
Chromatin remodeling, 466
Chromatofocusing, 38
Chromatographic distribution coefficient, 38
Chromatographic methods, characteristics of, 36t
Chromatography
  adsorption, 36–37
  affinity, 38–40
  in biomacromolecule purification, 34–40
  gel filtration/permeation, 35–36
Chromophores
  amide, 212
  perturbation of, 189–190
  role in protein absorption, 187–188
Chromosomal proteins, 140
Chromosomes, 140–141
  gene/physical maps of, 562
  telomers of, 77
Chymotrypsin, 366–368
Circular (cyclic) proteins, 129–130
Circular dichroism (CD)
  empirical applications of, 212–214
  main applications of, 213
Circular dichroism spectroscopy, 208–209. *See also* CD spectra
Circular permutation, 704
Circular proteins, 131t
cis interactions, 301
Class, Architecture, Topology, Homology (CATH) database, 608. *See also* CATH protein domain database
Cleavage agents, 97t
Cleland-form rate equation, 336
Cleland nomenclature, 339, 340t
Clone fingerprinting, 565
Clone libraries, 564
Cloning, DNA, 494–496
Cloning vectors, 494, 496
Closed β-α-β barrel structures, 135
Closed circular DNA duplex, topological transition of, 276
ClusPro Web site, 629
ClustalW algorithm, 523, 689, 697
Clustal X program, 523
Cluster analysis, 532–533
Clustering, categories of, 533
^{13}C-magnetic resonance (CMR), 198, 355
  monosaccharides and, 160–161
CM program, 285
Coatrahelicase, 454
Coding gap, translation over, 483
Coding sequence (CDS), 563
Codon bias detection, 576–577
Codons, 444–445
  redefinition of, 483
  redirection by frameshifting, 482
Codon usage database, 576
Coenzyme specificity, conversion of, 505
C—O ester bonds, 18
Cofactors, introduction at binding sites, 385–386
Coil-helix transition, in polypeptides, 273–274
Collagen, 129
Colligative properties, 46
Collisional activation process, 103–104
Collisionally activated dissociation (CAD), 104
Collisionally induced dissociation (CID), 104, 632–633
Collision quenching, 192
Color raster devices, 538
Column chromatography, 34
Combinatorial libraries, 241, 242
Combinatorial synthesis, 241–244, 584
Comparative genomics, 561
Comparative (homology) modeling (CM), 618, 619, 623–624
Complementarity, in nucleic acid interactions, 305–311
Complementarity-determining regions (CDRs), 303
Complex Carbohydrate Structure Database (CCSD), 662
Complex Carbohydrate Structure Data of SugaBase, 173
Complex formation, kinetic consequences of, 327
Compound libraries, 241
Computational approaches, categories of, 250
Computational genome annotation, 572
Computer Aided Spectrum Evaluation of Regular Polysaccharides (CASPER), 664
Computer-assisted molecular modeling, 253
Computer devices, 533–535
Computer graphics, of enzymatic reactions, 361
Computer graphic tools, 218
Computer languages, 536
Computers, types of, 533. *See also* Software
Computer system architecture, 534

Computer technology, 533–548
  languages and programming, 535–537
  molecular graphics, 537–540
Configurational isomerism, 9–10
Conformation(s)
  of biomacromolecules, 10–11
  comparing, 262
  in polysaccharide chains, 161–163
  in polysaccharide structures, 163–166
Conformational changes
  in lectin-carbohydrate recognition, 318
  tRNA-induced, 311
Conformational energetics, 143–144
Conformational entropy, 262–263
Conformational isomerism, 10–11
Conformational maps, 108–110
Conformational search, 253, 260, 261–262
Conformational space, searching, 260
Conformation prediction, side chain hydrophobicity/hydrophilicity and, 279
Conjugate gradient method, 257
Connection (loop) regions, 115–116
Conserved Residue Attributes (CORA), 136–137
Consortium for Functional Glycomics (CFG), 149, 553, 662
Constituent displacement, 43
Constitutional isomers, 11
Contiguous blocks (contigs), 564–565
Continuous DNA synthesis, 447
Contributing structures (contributors), 22–23
Controlled pore glass (CPG), 231
Control Unit (CU), 534–535
Cooperative binding system, 296
Cooperativity, 380
  in biomacromolecular interaction, 295–299
  diagnostic tests for, 299
Cooperativity coefficient, 271
Co-translational translocation, 489, 490
Cotton effects, 209, 210–211
Counter ions, 37
Coupling reactions, 225
  electrophilic site activation for, 229t
Couplings, 200–201
Covalent biomacromolecules, 12
Covalent bonds, 4–5
Covalent catalysis, 346–348, 393
Covalent glycan immobilization, 677–678
Covalent modifications, of enzymes, 374–375
Crabtree effect, 377–378
Critical Assessment of Predicted Interactions (CAPRI), 629
Critical Assessment of Techniques for Protein Structure Prediction (CASP), 619
Cross-linkers, 230
Cross-strand purine stacks, 87
Cross-validation procedure, 217–218
Cruciform DNA structures, 74–75
Crystalline properties, of starches, 164
Crystallographic R-factor, 218
Crystallographic study, of biomacromolecules, 216–219
Crystals, 214
  determination of X-ray diffraction patterns for, 216
CSA Web site, 330
C-terminal ladder sequencing, 101
C-terminal residue, determining, 98
C-type animal lectins, 313–314, 319
C-type cruciform formation, 75
CUTG Web site, 444, 563
3′,5′-Cyclic adenosine monophosphate (cAMP), 406–407. *See also* cAMP-response element binding protein (CREB)
Cyclic (rotation) symmetry, 137–139
Cyclic cascades, 375–377
Cyclophilin, 505
Cysteine, characteristics of, 20t
Cysteine proteases, 420
Cytosolic protein folding, molecular chaperones in, 494
Cytosolic protein tyrosine phosphatases, 416
Cytosolic ubiquitin (ATP-dependent) protein degradation, 430–433

D
DARWIN tool, 283, 284
Data analysis, in microarrays, 531–533
Database management system (DBMS), 548
Databases
  biochemical, 549–551
  components of, 549
  file formats for, 539
  interrogation of, 518
  retrieval of, 551–553
  searches of, 576
Data normalization process, 532
Dayhoff mutation data (MD) matrix, 519
Dayhoff PAM 001 matrix, 695
Dbcat Web site, 547
dbSNP Web site, 568. *See also* Single nucleotide polymorphisms (SNPs)
D-DNA, 69
Decarboxylases, 324
Decarboxylation, antibody catalyzed, 385
Decoding mechanisms, 479–481
Deconvolution procedure, 241
DeepView program, 627
Defined ordered DNA sequences (dosDNA), 73
Degree of polymerization (DP), 1, 148
Dehydratases, 324
Dehydration reaction, steps in, 221
Dehydrogenases (DHs), 323, 362, 363–364
De novo gene predictions, 575

Deoxynucleoside triphosphate (dpppN, dNTP) analogs, 60
Deoxynucleotides (dNTPs), 60
Deoxyribonucleic acid (DNA), 2. *See also* DNA entries
  A-form, 68, 72, 86, 213
  alternative structures of, 69–77
  annealing and denaturing, 275
  assembly of, 11
  B-form, 66–68, 70–73, 213
  as the carrier of genetic information, 442–443
  cruciform structures in, 74–75
  detecting functional sites in, 577–578
  dimeric binding sites in, 307–308
  duplication and genetic transmission of, 443–444
  electrophoretic separation of, 42
  generalizations concerning, 559–560
  handedness of, 72
  melting temperature of, 51–52
  secondary structure of, 63–77
  sequence-specific recognition of, 306
  slipped, mispaired, 74
  supercoiling and tertiary structure of, 77–80
  symmetry elements in, 73
  tetraplex, 77
  triplex, 73, 75–77
  unwinding, 310
  use in nanosciences, 91
  Z-form, 69
Deoxyribozymes, 325
Dephosphorylation, in signaling, 414–417
Diabodies, 511
Diagnostic tests, for cooperativity, 299
Diamino carboxylic acid, in peptide synthesis, 232–233
Diastereisomers, 9
Dichroism, 195. *See also* Circular dichroism spectroscopy
  IR, 196, 197
Dideoxy DNA sequencing, 58–59, 60–61
Dideoxynucleoside triphosphates (dd-pppNs, ddNTPs), 60
Dideoxynucleotide (ddNTP), 60
Dielectric constant, molecular, 255
Differential display PCR (DD-PCR), 583–584. *See also* Polymerase chain reaction (PCR)
Diffraction, 214
  radiation scattering via, 183
Diffraction pattern, 214
Dihedral angles, in conformational search, 261
Dihedral symmetry, 139
Dimensionality reduction methods, 532
Dimeric binding sites, 307–308
Dinucleotides, 241
Dipolar ions, 21
Dipole–dipole interactions, 255
DIP Web site, 300
Direct condensation, for biomolecular synthesis, 229t
Direct condensation coupling reaction, 225
Discontinuous DNA synthesis, 447
Disease, altered glycoforms associated with, 660t
Distance-based phylogenetic method, 690–691
Distance Mapping Web site, 663
Distributive enzymes, 436
Disulfide bonds, proteins with, 124
D-loop (displacement loop) replication, 448
DnaA protein, 453
DNA array hybridization, 583. *See also* Deoxyribonucleic acid (DNA)
DNA-binding domain (BD), 636
DNA chain structures, 58t
DNA changes, interactions governing, 69–70
DNA chips. *See* DNA microarrays (DNA chips)
DNA cloning, 494–496
DNA Data Bank of Japan (DDBJ), 57, 548, 570–571, 601
DNA-directed DNA polymerase activity, 455
DNA duplex structures, 70–71t
  closed circular, 276
DNA excision repair, 459–460
DNA fragments, electrophoretic mobility of, 42
DNA gyrases, 453
DNA libraries, 564–566
DNA ligases, 450–451
  properties of, 452t
DNA methylation, 456–457
DNA microarrays (DNA chips), 525, 578–582. *See also* Microarray entries
  mutation monitoring and, 568
  PCR products for, 528
DNA migration, 42
DNA photolyases, 459
DNA polymerase activity
  DNA-directed, 455
  RNA-directed, 455
DNA polymerases, 60, 307, 448–450
DNA polymorphism, 66–69
DNA-protein interactions, 305–309
  principles guiding site-specific recognition in, 308
DNA rearrangement, 560
DNA recombinant technology, 504
DNA repair, 458–461
DNA replication, 444, 445–461, 453
  enzymology in, 448–455
  fidelity of, 447
  general features of, 445–448
DNA restriction endonucleases, 457
DNA secondary structures, helical parameters for, 67t
DNA sequences, 563
  defined ordered, 73
DNA sequencing
  breakthroughs in, 57–58
  commercial dyes for, 62–63
  protocols for, 58–61
DNA sequencing technology, automated, 62–63
DNA shuffling, 498
DNA strands, number and orientation of, 73

DNA structure, double-stranded, 66
DNA synthesis
  polarity of, 447
  SOS translesion, 461
DNA template, binding of RNA polymerase to, 462
DNA topoisomers, 77–80
Docking protocols, 283
DOCK tool, 284
DockVision tool, 284
Domain duplication, 703
Domain families, 703
Domain names, 543t
Domain structures, in proteins, 119–121
Dot matrix representation, in sequence comparison, 518–519
Dotplot, 518–519
DOTTER program, 519
Double-bond diastereomers, 9
DOUBLESCAN, 573–575
Double-strand breaks, induction of, 701
Double-stranded DNA (dsDNA), 66, 453–454
  helical repeat of, 70
  helix-coil transition in, 274–275
Double-stranded genomes, 558
Double-stranded RNA (dsRNA), 85
Drude equation, 210
Dynamic Molecules Web site, 663
Dynamic programming algorithm, 536
Dynamics simulation, 253

E
Eclipsing strain, 11
Edman degradation, 99, 100, 101
Edman sequanators, 100–101
Effector enzymes, 406–408
Effectors, control of enzyme catalytic activity by, 377–380
EF-hand motif, 411
EF-Tu · GDP, 479–480
Electrofocusing (EF), 42–43
Electron density map, 217
Electronic spectra, 185–187
Electronic transitions, 208
Electron transfer, 363
  flavoenzyme catalyzed, 364
Electrophoresis
  in biomacromolecule purification, 40–44
  capillary, 42
  high-resolution polyacrylamide gel, 57
  polyacrylamide gel, 41
  two-dimensional, 43–44
  zone, 41, 43
Electrophoretic mobility (U), 40
Electrospray ionization (ESI), 48, 50, 61. *See also* ESI-MS
Electrostatic catalysis, 346
Electrostatic interactions, 6
Electrostatic stabilization, 328
Eliminases, 156
Ellipticity, 209, 211
Elongation cycle, 478, 479
EMBL nucleotide sequence database, 552, 570. *See also* European Molecular Biology network (EMBnet); Translated EMBL (TrEMBL)
Emission, of radiation, 185
Empirical force field (EFF), 252
Enantiomers, 9
Endoglycanases, 422–423
Endoglycosidases, 53, 156, 175, 176t
Endonucleases, 425
Endopeptidases, 426
Endoplasmic reticulum (ER) membrane, 489
Energy-based modeling, 623
Energy calculation/minimization, 252, 254–256. *See also* Energy minimization (EMin); Energy minimum
Energy metabolism, 398
Energy minimization (EMin), 256–258
  example of, 265–267t
Energy minimum, 249
Enhancer promoter interaction, 468
Enhancers, 467
Enthalpy change, in hydrophobic interaction, 7
Entrez molecular biology database and retrieval system, 551, 601
Entrez Web site, 523, 547
Entropic effect, 344–345
Entropy change, in hydrophobic interaction, 7
Environmental factors, in enzymatic reactions, 341–344
Enzymatic chain termination DNA sequencing, 58–59, 60–61
Enzymatic degradation, 100
Enzymatic hydrolysis, 96
Enzymatic nucleophiles, 347–348
Enzymatic photoreactivation, of pyrimidine dimers, 459
Enzymatic reactions
  environmental factors in, 341–344
  pH effect in, 343–344
  in the presence of inhibitors, 341–343
  rates of, 333
  temperature effect in, 343
Enzymatic sequence analysis
  chemical methods for, 99–100
  enzymatic methods for, 100–101
Enzyme active sites, 280, 326, 330–331
Enzyme catalysis
  control by effectors, 377–380
  factors in, 333–333
  steady-state kinetic treatment of, 336–337
Enzyme-catalyzed reactions
  transition state structures of, 357
  transition state theory for, 356
Enzyme cofactors, 325
Enzyme Commission (EC) nomenclature, 323
Enzyme Commission numbers, 324
Enzyme compartmentation/solubility, 374
Enzyme concentration, 374
Enzyme engineering, by site-directed mutagenesis, 501–504

Enzyme kinetic data, treatments of, 334
Enzyme kinetics, 333–344
  approaches to solving, 334
Enzyme mechanisms, 344–374
  active site studies of, 349–356
  case studies of, 361–374
  structure-activity relationships in, 357–360
  transition state studies of, 356–357
  X-ray crystallographic studies of, 361
Enzyme prodrug therapy, antibody-directed, 510
Enzyme-product (EP) complex, 336
Enzyme reactions
  Cleland nomenclature for, 339
  reversible inhibition patterns of, 342t
Enzyme regulation, 374–383
  allosteric regulation, 381–383
  covalent modifications and cascade effect, 374–377
  effector control of enzyme catalytic activity, 377–380
  elements of, 374
Enzyme resource sites, 324t
Enzymes
  as biocatalysts, 323–324
  characteristics of, 325–333
  distributive versus processive, 436
  interconversion of, 375
  nucleophilic groups in, 346t
  plasticity and flexibility of, 326
  redesigning, 504
  specificity of, 326, 328–329
  stereospecificity of, 348–349
  subcellular localization of, 33t
Enzyme Structure Database, 616
Enzyme-substrate (EA) complex, 336
ENZYME Web site, 324, 610
Enzymology, in DNA replication, 448–455
Epidermal growth factor (EGF), 417
Epimers, 9
ε-amino groups, acetylation of, 484
Equilibria, multiple, 291–295
Equilibrium binding processes, 290
Equilibrium dialysis, 290
Equilibrium measurement, in binding studies, 292
Equilibrium reactions, 323
Erythropoietin receptors, 419
Escherichia coli translation, protein factors of, 475t
E selectin, 314
ESI-MS, 632. *See also* Electrospray ionization (ESI)
EST databases, 566t, 576
Ethidium bromide, binding to DNA, 310
Eukaryotes, protein synthesis initiation in, 476–477
Eukaryotic DNA, 559–560
Eukaryotic DNA polymerases, 450
  properties of, 451
Eukaryotic genes, 564
  regulatory influences on, 467
Eukaryotic mRNAs, 467, 470
  purification of, 39
Eukaryotic peptide elongation, 479
Eukaryotic ribosomes, 141t
Eukaryotic RNA transcription, 463–464
Eukaryotics, gene regulation in, 457
Eukaryotic signal peptidases, 490
Eukaryotic transcription regulation, 466–467
Eukaryotic transcripts, posttranscriptional processing/modification of, 469
European Bioinformatics Institute (EBI), 56–57, 552, 570
European Molecular Biology network (EMBnet), 547. *See also* EMBL nucleotide sequence database
E-value, 524–525
Evolution
  of biosequences, 697–701
  role of mutation in, 680–682
Evolutionary change
  rate of, 682–685
  regulation of, 699–701
Evolutionary genes, 681
Evolutionary tree, 685–686
Excitation spectrum, 190
Exoglycanases, 422–423
Exoglycosidases, 156
Exons, spliced, 470–471
Exonucleases, 425
Exopeptidases, 426, 427
Expanded genetic code, 482
Expert Protein Analysis System (ExPASy), 595, 610–612
Explicit solvent model, 625–626
Expressed protein ligation (EPL), 245, 246, 675
Expressed sequence tags (ESTs), 528, 565–566
Expression profiling, 578–582
Expression proteomics, 594, 595
Extended site multivalency, in lectin-carbohydrate recognition, 317–318
Extracellular signal regulated protein kinase (ERK), 418
Extrathermodynamic relationships, 358, 359–360
Extrinsic pathway, 421
EzCatDB Web site, 344

F
Fas-FasL system, 422
FASTA program, 521–522, 523, 599
Fast atom bombardment (FAB), 48
Feedback control, 378, 379t
Fibrous proteins, 94, 128–129
File Transfer Protocol (FTP), 545
First messengers, 400, 401t
Fitch–Margoliash method, 691
Flat-file databases, 551
Flavin coenzymes, 362
Flavo-dehydrogenase substrates, 366
Flavoenzymes, 362
Flavosemiquinone, 363
Fletcher–Reeves approach, 257
FlexX tool, 284
Flicker-server Web site, 631
FLOG tool, 284

Fluorescein isothiocyanate (FITC), 529–530
Fluorescence-based dideoxynucleotide sequencing chemistries, 62
Fluorescence detection, 192
Fluorescence efficiency, 190
Fluorescence spectroscopy, 190–192
  applications of, 191
Fluorescent probe (FP), 499
Fluorescent spectra, rules for interpreting, 191
Fluorophores, 190
  characteristics of, 191t
FlyBase Web site, 605
fMet-$\text{tRNA}_{\text{f}}^{\text{Met}}$ initiator, 475–476
Fmoc/tBu protecting groups scheme, 235
Folded proteins, molecular dimension of, 51
Folded structures, accessible surface of, 130–133
Folding behaviors, RNA versus protein, 491
Fold recognition (FR), 618, 619, 623
Folds, protein binding and, 126–127
Force field (FF), 252, 254, 256
Four-helix bundle structures, 120
  classes of, 122–123
Fractional analysis, 155
Fractional saturation, 291, 296–297
Fragment ions, 50
Frameshifting, 482
Frameshift mutagenesis, spontaneous, 74
Framework protein folding model, 493
Free-energy change, 289
Free energy of activation, 322–323
Free energy of supercoiling, 80
Free induction decay (FID), 202
Freeware, 540
  biocomputing, 537t
Frequency-correlation, 196
Frequency-sweep method, 199
Frictional ratio, 51
FTDOCK tool, 283, 284
FUGUE Web site, 626–627
Functional groups, protection and deprotection of, 221–223
Functional site residues, mapping of, 281
Functional sites, 614–616
Function-based (functional) microarrays, 639
Furanose rings, 161
  five-member, 24–27

G
Gal (galactose), ligand discrimination of, 318–320
Galactose packing interactions, 319
Galectins, 314
γ-carboxyglutamate (Gla), 484
Garnier, Osguthorpe, and Robson (GOR) method, 278–279
GAT superfamily, 707–708
Gaussian type orbitals (GTO), 250
G·C base pairing, 275
GC box, 467
Gel electrophoresis, 41
  two-dimensional, 629–631
Gel filtration/permeation chromatography, 35–36
Gel pore size, effect on charged biomacromolecule separation, 41–42
Gels, pectin, 165
GenBank, 56, 569–570
Gene cataloguing, approaches to, 587
Gene duplication, 702
Gene expression, 578–587
  databases, 582
  role of supercoiling in, 79–80
  serial analysis of, 584
Gene identification
  approaches to, 571–578
  sequence analysis techniques for, 572–575
Gene identification programs, Web-based, 575t
GeneID Web site, 575
Gene linkage maps, 562
Gene locus, 560
Gene mapping, 561–563
Gene ontology (GO), 553–554
Gene Ontology Annotation (GOA), 554
Genetic algorithm, 625
Genetic codes, 444
  expanded, 482
Genetic Codes Web site, 563
Genetic information
  control of expression of, 306
  faithful translation of, 372
  horizontal transfer of, 701
  transmission of, 442–445
Genetic mapping, 561–562
Genetic variation, 567–568
  optimal, 681
Genie Web site, 575
Genome analysis, computer programs for, 563
Genome databases
  general, 561t
  mining of, 572
Genome informatics, 568–571
Genome Project. *See* Human Genome Project (HGP)
Genome project databases, 588–589t
Genome/proteome databases, integrated, 613–614
Genomes
  diversity of, 558
  organization of, 560
Genomic analysis servers, 574t
Genomic libraries, 564
Genomics, 515, 558–593. *See also* Gene expression
  approaches to gene identification, 571–578
  genome features and organization, 558–568
GenPept format, 601
GenScan Web site, 575
Geometric isomerism, 9
German Cancer Research Center (DKFZ), 548
Global forces, 305
Global metabolite profiling, 648–649
Global pair-wise alignment, 520–521
Global sequence alignment, 687–688
Globin fold structure, 120

Globular proteins, 94
  secondary structures of, 116–117
  supersecondary structures in, 117–118
Glucagon structure, geometric optimization of, 265–267t
Glucan secondary structures, 162
D-Glucose (D-Glc), 436
  biosynthetic conversion of, 437
Glutamic acid, characteristics of, 20t
Glutamine (Glu), 484
  characteristics of, 19t
Glutathine reductase (GR) catalysis, 364–366
Glycan affinity chromatography, lectins used in, 40t
Glycan analysis, 663–665
  online servers for, 664t
Glycan Binding Proteins (GBP) database, 665
Glycan combinatorial structures, 657
Glycan Database, 553, 662
Glycan formation, human genetic defects in, 667t
Glycan linkage, symbolic representation of, 149–150
GlycanMass tool, 612, 663
Glycan–protein interactions, 661
Glycans, 13, 147. *See also* Polysaccharides
  as biocatalysts, 323
  constituents of, 23–27
  covalent immobilization of, 677–678
  glycoprotein-associated, 167–177
  hydrolysis versus phosphorolysis of, 422–424
  information related to, 173–174
  linear code for, 148–153
  NMR of, 207–208
  nomenclature and representation of, 655–657
  non-covalent immobilization of, 676–677
  release from glycopeptide/glycoproteins, 175
  self-splicing, 325
Glycan signature (fingerprint), 177
Glycan structure(s), 661–663
  representation of, 148
Glycan structure analysis, 674
  enzymes used in, 157t
  facilitation of, 176
N-Glycan synthesis, 670–671
Glycation, 486–487
Glycine, characteristics of, 19t
Glycoanalysis, online utilities for, 666t
O-Glycobase Web site, 662–663
Glycobiology, 167–177, 655–657
Glycochips, 675–678
Glycoconjugates/glycosides, linear codes for, 153
Glycoenzyme-based proteoglycomics, 670–672
Glycoforms, 167–168, 657–659
GlycoFragments Web site, 665
Glycogen, 164–165
Glycogenes, 666
Glycogen phosphorolysis, 425
Glycogen phosphorylase, 381–383, 424
  regulation by phosphorylation/dephosphorylation, 376
Glycogen synthase, regulation by phosphorylation/dephosphorylation, 376
Glycolysis, inhibition by respiration, 377
Glycolytic genes, 667–668
GlycoMaps DB Web site, 662, 664
Glycomes, distinguishing features of, 660–661
Glycomic databases/servers, 661–666
Glycomics, 515, 655–679
  chemoglycomic approaches to, 674–678
  features of, 655–661
  genetic approaches to, 666–668
  glycoprotein syntheses in, 674–675
  objectives of, 660
  proteoglycomic approaches to, 668–674
  research in, 13
Glycomic tools, 662
GlycoMod tool, 612, 663
Glycopeptide bond synthesis, linkages and enzymes involved in, 671t
Glycopeptide/glycoproteins, release of glycans from, 175
Glycopeptides, 167
Glycoproteins, 167–168. *See also* Neoglycoproteins
  as biomarkers and therapeutic targets, 657
  biosynthesis of, 437–442
  glycosidic, 657
  linear codes for, 153
  N-linked glycans of, 172
  outer chains of, 173
  recombinant, 673–674
  removal of glycosides from, 53
  structural analysis of, 174–177
Glycoprotein syntheses, in glycomics, 674–675
Glycosciences Web, 553, 656
Glycose cyclic oxygen atoms, as hydrogen-bond acceptors, 316
Glycose residues, attachment of, 670
Glycoses, 23, 24, 173
Glycosidase-catalyzed reactions, 239
Glycosidases, 423–424, 670–671
Glycosides, removal from glycoproteins, 53
Glycoside synthesis
  orthogonal protection in, 221
  solid-phase, 238–241
Glycosidic cleavages, 158
Glycosidic glycoproteins, 657
Glycosidic hydrolases, 156
Glycosidic linkage, 161, 166
Glycosylated sites, identifying, 174–175
Glycosylation
  mucin-type, 173
  of proteins, 665–666
  versus glycation, 486–487

Glycosylation Pathways database, 665
Glycosylation sites, characterization of, 668–670
Glycosylphosphatidylinositol (GPI), 169
Glycosylphosphatidylinositol membrane anchor, 440–442. *See also* GPI-anchored proteins
Glycosylphosphatidylinositol protein, biosynthesis of, 442
Glycosyl transferases, 239, 438. *See also* Glycotransferases
Glycosynthetase, 505
Glycosynthetic genes, 667–668
Glycotransferases, 671
GlycoWord Web site, 661
GO Consortium, 554
GOLD tool, 284
Golgi apparatus, 439
Google, 545
GPCRDB Web site, 404. *See also* G-protein-coupled receptors (GPCRs); PRED-GPCR Web site
gpDB Web site, 404
GPI-anchored proteins, in mammalian cells, 173. *See also* Glycosylphosphatidylinositol entries
G-protein-coupled receptors (GPCRs), 400. *See also* GPCRDB Web site; PRED-GPCR Web site
G-proteins. *See* GTP-binding proteins (G-proteins)
Graphics programs, 537–540
  molecular, 264t
Greek key barrels, 120, 121
Ground state, 184
Group I introns, 388–389
Group II introns, 389–390
Group transferases, 324
GTPase, 405–406
GTP-binding proteins (G-proteins), 400, 403–406
  events regulated by, 404
  GTPase activity of, 405
GTP hydrolysis, 481
G-U wobble pairs, 281, 282

H
Hairpin β motif, 118
Hairpin ribozyme, 391, 393
Haldanes' relationship, 336
Haloenzymes, 391
Hamiltonian (H), 259
Hammerhead RNA, 390–391
Hammett equation, 358
Hansch equation, 360
Haptens, design of, 386
H-chain gene assembly, 506
HDV ribozyme, 393–394. *See also* Hepatitis delta virus (HDV) RNA
Heat-shock proteins (HSPs), 494
Heavy (H) chains, 300–301
hect domain family, 432
Helical grooves, 68
Helical junctions, 72–73
Helical stacking, 81
Helical structures, length of, 113
Helical symmetry, 139
Helical transitions
  in ccDNA, 276
  in nucleic acids, 274–275
Helical wheel, 113
Helicases, 453–454
Helices. *See also* Helix entries
  role in recognition, 306–307
  triple, 75
Helix axis, deformation of, 72
Helix diameter, 65
Helix formation, spectral changes accompanying, 204
Helix pitch, 65
Helix sense, 65
Helix-turn-helix (αα) motif, 117, 308–309
Hemiacetals, 23
Hemiketals, 23
Hepatitis delta virus (HDV) RNA, 391. *See also* HDV ribozyme
Heptasaccharide, 170
Heteroduplexes, 581
Heterogeneous nuclear RNA (hnRNA), 85
Heteroglycans, 3
  linear codes for, 153
Heterologous association, 137
Heteropolymers, sequential, 1
Heteropolynucleotides, 3
Heteropolysaccharides, 148, 149
Heterotrimeric G-protein, 404–405
Heterotropic transition, 382
Heuristic techniques, 692
Hexopyranose residues, periodate oxidation of, 156t
Hidden Markow models (HMMs), 605
High methyl pectins (HM-pectins), 165
High performance liquid chromatography (HPLC), 28, 39
High-resolution PMR, 206
High-resolution polyacrylamide gel electrophoresis (hrPAGE), 57. *See also* Polyacrylamide gel electrophoresis (PAGE)
High scoring pairs (HSPs), 522
High-throughput protein crystallography (HTPC), 635–636
High-throughput screening (HTS) techniques, 676
Hill equation, 294
Hill's coefficient, 294, 299
Hill's plot, 295, 299
Histidine, 303
  characteristics of, 20t
  in peptide synthesis, 232
Histone Database, 140
Histones, 140
  acetylation of, 484
HIV-1 reverse transcriptase, 455–456
hnRNPs, 470
Hofmeister series, 36
Holliday junction of genetic recombination, 72–73
Homoglucan, biosynthesis of, 438
Homoglycans, biosynthesis of, 436
Homologous amino acids, 644
Homologous sequences, 517
  detection of, 598
Homologous structures
  catalytic site divergence in, 707–708
  homology modeling based on, 277
Homologues, searching for, 522

Homology modeling, 277, 618, 619, 623–624
Homopolynucleotides, 2
Homopolypeptides, random-coil, 211–212
Homopolysaccharides, 148, 149
  conformations of, 163
Homopurine-homopyrimidine DNA region, 76
Homotropic interactions, 378–380
  models of, 296–298
Homotropic transition, 382
Hoogsteen base-pairing, 76
Horizontal gene transfer, 680
HotBot, 545
Housekeeping genes, 467
Hsp70 molecular chaperone system, 494, 495
Human gene expression databases, 575
Human genome, 558
Human genome information, medical application of, 589–590
Human Genome Project (HGP), 587–590
Hybridization, in DNA microarray analysis, 580t
Hydration/solvation, in biomacromolecule simulation, 263
Hydrazide covalent conjugation, 668
Hydrazinolysis, 53, 175
Hydrogen bond donor/acceptor sites, 307
Hydrogen bonding, 70, 255–256. *See also* Hydrogen bonds
  in lectin interactions, 315
  between protein and DNA atoms, 305
  via 2′-hydroxyls, 86
Hydrogen bonds, 5, 6–7. *See also* Hydrogen bonding
  in antibody-antigen interactions, 304–305
  in the DNA double helix, 89
  glycose functionalities that form, 316
  lengths and strengths of, 7
  reverse Hoogsteen, 76
Hydrolases, 324, 366–368
Hydrolysis
  arginase catalyzed, 369
  chymotrypsin catalyzed, 367
  of glycans, 422–424
  to monosaccharides, 154
Hydrophobic collapse protein folding model, 493
Hydrophobic interactions, 7
Hydrophobicity, 360
Hydroxyapatite, 36–37
Hydroxy protecting groups, 224t
HyperChem, geometric optimization of glucagon structure using, 265–267t
Hyperchromic effect, 51
HyperText Markup Language (HTML), 544

I
Icosahedral symmetry, 139
"Identity swap" experiment, 311
Ig genes, 506. *See also* Immunoglobulin entries
IMGT Web site, 301
Imidazole protecting groups, 228t
Imino proton resonances, 206
Immobilization approach, to monosaccharide sequence determination, 177
Immobilized pH gradients (IPG strips), 43–44
Immune response, diversity of, 300
Immunoglobulin fold, 301
Immunoglobulin G (IgG), 300, 301
  L chains of, 302
Immunoglobulin genes, assembly of, 506
Immunoglobulin receptors, 419
Immunoglobulins (Igs), 325
  affinity and specificity of, 386
  homology regions of, 301
  properties of, 300t
Immunological databases, 506
Immunological response, cloning into *Escherichia coli*, 386
Immuno-PCR (IPCR), 498. *See also* Polymerase chain reaction (PCR)
iMolTalk Web site, 608
Implicit solvent model, 626
Induced shift, 200
Inducers, 466
Inference from trapped intermediates approach, 357
Informatics, 515, 548–553. *See also* Bioinformatics; Genome informatics; Proteome informatics
Infrared spectroscopy, 193–197, 352–353. *See also* IR entries
  biochemical applications of, 195–197
Inhibition kinetics, 341–343
Inhibitor/activator, effect on allosteric enzymes, 380
Inositol triphosphate receptors, 410
Insertions/deletions (InDels), 704
IntAct project, 300
Integr8 portal, 613
Integrated databases, 613–614
Intein-mediated protein legation (IPL), 245
Inteins, expressed protein ligation using, 245
Interacting protein databases, 300
Interaction constants, measurement of, 290
Interactive docking, 283
Interatomic shielding, 200
Intercalation agent, binding to supercoiled DNA, 309–310
Inter-enzyme complexes, X-ray crystallographic determination of, 330–331
Interferon, 81
Interleukin-2 receptor (IL-2R), 419
Intermediary metabolism, 398
Intermediates, multiple, 328
Internal RNA loop motifs, 87
International Nucleotide Sequence Database Collaboration (INSDC), 568–569
International Union of Pure and Applied Chemistry (IUPAC) nomenclature, 11. *See also* IUPAC entries
Internet, versus intranet, 546

Internet browsers, 545
Internet Protocol (IP), 540–541
Internet resources, 540–546
  for biological sciences, 546–548
Internet Society (ISOC), 546
InterPro Web site, 602, 613
Intervening sequences (IVSs), 387, 390, 563
Intracellular communications, 399–400
Intramolecular potential energy, 251
Intramolecular transfer reaction, facilitation by multienzyme complexes, 332
Intramolecular triplex structures (H-DNA), 76, 77
Intranet, versus Internet, 546
Intrinsic pathway, 421
Intrinsic shift, 200
Introns, 387
  Group I, 388–389
  Group II, 389–390
Intron splice sites, detecting, 577
In vacuo solvent effects, 249
Inverted repeat sequences, 74
Inverting glycosidases, 424
In vitro protein folding pathway, 492–494
Ion channels, oligomeric, 402–403
Ion-exchange chromatography, 37–38
Ionic interaction (salt bridge), 305
Ionogenic exchangers, 37t
IR dichroism, 196, 197
IR frequencies, characteristic group, 194t
IR measurements, 195
Irreversible inhibitor, 341
IR spectra, 193, 195, 197. *See also* Infrared spectroscopy
ISIS Draw, 263–264, 538
Isoelectric focusing (IEF), 42–43, 629
Isoelectric pH, of proteins, 52–53
Isoelectric point, 21
Isoleucine, 18
  characteristics of, 19t
Isologous association, 137
Isomerases, 324
Isomerism, 8–11
Isomerization, thiol-disulfide, 487
Isomorphous crystals
  determination of X-ray diffraction patterns for, 216
  structure factors and heavy atoms for, 216–217
Isomorphous heavy-atom derivatives, preparation of, 216
Isopycnic centrifugation, 52
Isotope-coded glycosylation-site-specific tagging (IGOT) process, 668, 669
Isozymes, 326
I-Type lectins, 314
IUBMB Web site, 661. *See also* IUPAC-IUBMB Nomenclature of Carbohydrates
IUPAC carbohydrate nomenclature, 148–149. *See also* International Union of Pure and Applied Chemistry (IUPAC) nomenclature
IUPAC-IUBMB Nomenclature of Carbohydrates, 655. *See also* Linear Notation for Unique description for Carbohydrate Sequences (LINUCS)

J
J correlated spectroscopy (COSY), 203
Jelly roll barrels, 120–121
J splittings, 203

K
KABAT Web site, 301
KEGG Carbohydrate Matcher (KCaM), 664. *See also* Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway Database
Keto-enol equilibrium, 15, 16
Keyword search, 523
KFERQ proteins, 430
KineMage, 218, 264
Kinetic energy, 251, 259
Kinetic enzyme reaction mechanism studies, 344
Kinetic isotope effects approach, 357
Kinetic parameters, implications of, 335
Kinetic PCR, 499. *See also* Polymerase chain reaction (PCR)
Kinetic processes, in cruciform DNA formation, 75
Kinetic proofreading, 479–480
Kinetics, nonlinear, 339–341
Kinetic studies, of GR catalysis, 365
King and Altman method, 336–337
KinTekSim Web site, 344
Klenow fragment of DNA polymerase I (KF-Pol I), 450
Klotz equation, 293
Klotz plot, 295
Knots, DNA molecule, 80
Knowledge-based modeling, 623
Knowledge-based structural prediction, 277
Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway Database, 398. *See also* KEGG Carbohydrate Matcher (KCaM)

L
Ladder sequencing, 60
Langevin dynamics, 260–261
Lattice points, in crystals, 214
Laue's equations, 215
L-chain gene assembly, 506
L chains, IgG, 302
Lectin affinity capture, 668
Lectin-carbohydrate recognition
  general, 315–318
  ligand discrimination in, 318–320
Lectin families, 313t
Lectin-Gal complex structures, 316
Lectins, 670
  applications of, 312
  binding sites of, 317–318

- in carbohydrate-lectin interaction, 312–320
- classification and structures of, 312–315
- differential binding of glycoses to, 318

Legume lectins, 312
Lennard–Jones potential, 6, 8, 255
Leucine, characteristics of, 19t
Leucine zipper, 309
Libraries
- combinatorial, 241, 242
- compound, 241

Ligand-based approach, to predicting functional sites, 280–281
Ligand discrimination, in lectin-carbohydrate recognition, 318–320
Ligand-induced conformational changes, 295
Ligand-receptor interactions, 282
- allosterism and cooperativity in, 295

Ligands
- for affinity chromatographic protein purification, 39t
- binding to receptor biomacromolecules, 282
- systematic evolution of, 584–587

Ligases, 324, 372–374
Light amplification by stimulated emission of radiation (LASER), 185. *See also* MALDI-TOF MS
Light (L) chains, 300
Light scattering, in determining molecular weight, 47
LIGIN tool, 285
LiGraph tool, 662
Linear amplification, 496
Linear β-α-β proteins, 135
LinearCode, 150–153, 657
Linear glycans, linear codes for, 151–152
Linear inhibition, 342–343
Linear molecules, 13
Linear Notation for Unique description for Carbohydrate Sequences (LINUCS), 150, 656. *See also* IUPAC entries
Line drawings, computer, 537–540
Linkers
- in activity-based probes, 642
- on polymeric supports, 230

Linking number, 78, 80
Lipid-linked core precursor, synthesis of, 439
Lipopolysaccharides, linear codes for, 153
Liquid chromatographic methods, 35–40
Liquid chromatography (LC), 34
Liquid–liquid partition chromatography, 37
Liquid–solid adsorption chromatography, 36–37
List servers, 542–543
Liver glycogen, 164
Local alignment, 521–522
Local forces, 305
Locally disrupted base pairing, forms of, 73–74
Local sequence alignment, 688
Long interspersed nuclear element (LINE) sequences, 698
LongSAGE, 584
Low energy conformations, determining, 261–262
Low methyl pectins (LM-pectins), 165
L selectin, 314
LUDI tool, 285
Lyases, 156, 324
Lysine, characteristics of, 20t
Lysosomal protein degradation pathway, 430
Lysozyme catalysis, 345
Lysozyme structure file, 541

M

Macromolecular conformations by symbolic programming (MC-SYM) approach, 282
Macromolecular structure, time-average conformation of, 264–267
Macromolecules. *See also* Biomacromolecular entries; Biomolecular entries; Molecular entries
- conformational entropy of, 262–263
- conformations of, 255
- hydrating/solvating, 263
- structural hierarchy of, 55–56

MALDI-TOF MS, 632, 674. *See also* Matrix-assisted laser desorption/ionization (MALDI); Matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometers
MALIGN Web site, 689
Mammalian cells, GPI-anchored proteins in, 173
Mammalian DNA ligases, properties of, 452t
Mammalian genomic imprinting, methylation in, 457
Man/Glc (mannose/glucose), ligand discrimination of, 318–320
MAP kinase-ERK kinase (MEK), 418. *See also* Mitogen activated protein kinases (MAP kinases)
Maskless array synthesizer (MAS), 529
Mass spectrometers, components of, 48
Mass spectrometric analysis, 61, 158
- of posttranslational modification, 634–635

Mass spectrometry (MS). *See also* MALDI-TOF MS; Tandem mass spectrometry (MS-MS)
- in determining molecular weight, 48–50
- in protein chemistry, 101–103
- proteome analysis by, 631–634
- sequencing via, 104

Matrix-assisted laser desorption/ionization (MALDI), 48, 50, 61
Matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometers, 102–103. *See also* MALDI-TOF MS

Maxam–Gilbert sequencing method, 59
Maximal-scoring segment pairs (MSPs), 524
Maximal separation rule, 348
Maximum likelihood method, 691–692
Maximum-match matrix, 521t
MCDOCK tool, 285
Mean-square (ms) displacement, 272
Mechanism-based probes, 642
Medicine, application of human genome information in, 589–590
Meiotic recombination, 560
Melting temperature, of DNA, 51–52
Membrane-bound proteins, 124
Membrane folds, 128
Membrane proteins, 128, 136
Membrane spanning sequences, 403
Messenger RNA (mRNA), 81–82, 84, 461. *See also* Eukaryotic mRNAs; mRNA entries
  folding of, 492
  methylated, 471
  truncated, 483
Metabolic oligosaccharide engineering, 672–673
Metabolism, 398–400
  purposes of, 399
Metabolite profiling, 648–649
Metabolite shuttles, 374
Metabolome, 647–649
  analysis of, 648, 649t
  profiling, 648
Metabolomics, 515, 647–648, 649
Metal binding RNA motifs, 88
Metal cofactors, introduction at binding sites, 385
Metal ion catalysis, 346
  versus acid–base catalysis, 392–393
Metal ions
  in hydrolase catalyses, 368
  in lectin-carbohydrate recognition, 317
  in RNA, 86
Metalloproteases, 427
Met-containing peptides, 232. *See also* Methionine; Met-$tRNA_i^{Met}$ initiator
Methionine, characteristics of, 20t. *See also* Met-containing peptides; Met-$tRNA_i^{Met}$ initiator
Methylation, 484
  regulation of genetic change by, 700
Methylation analysis, 155
Methylnitropiperonyloxycarbonyl (MeNPOC) group, 579
Methyltransferase, 456
Metropolis method, 261
Met-$tRNA_i^{Met}$ initiator, 476. *See also* Met-containing peptides; Methionine
Michaelis–Menten constant, 335
Michaelis–Menten kinetic behavior, 338
Microarray analysis, 530
Microarray complexity, 581
Microarray databases/servers, 582t
Microarray detection, 530–531
Microarray manufacturing technology, 529
Microarray probes, 529–530
Microarrays, 525–533. *See also* DNA microarrays (DNA chips)
  biochemical reactions of, 530
  data analysis in, 531–533
  function-based, 639
  genotyping applications of, 581
Microarray surface
  affixing proteins to, 639–640
  preparation of, 525–527
Microarray targets, 528–529
  approaches to preparing, 528
  attachment of, 527
Microarray technology, advances in, 578
Microenvironments, 328
Microheterogeneity, 168, 169, 659
MicroRNA (miRNA), 85, 469
Microscopic ligand association constant, 292
Microspray technique, 102
Miller indices, 215, 216
Minimal motion rule, 348–349
Minimal number rule, 348
Minor bases, 16–18
miRISC, 469
Mitogen activated protein kinases (MAP kinases), 416, 418
Mixed α helix/β strand packings, 123–124
Mixed β-sheets, 114
Mixture synthesis, 242–244
MMDB Web site, 610
ModBase Web site, 627
Modulator enzymes, 414–416
Moffitt equation, 210
MolD simulation, 260, 267. *See also* Molecular dynamics (MolD)
  of biomacromolecules, 259–260
Molecular chaperones, in cytosolic protein folding, 494
Molecular cloning, 494
Molecular docking, 282–285
  strategies for, 283
  tools (programs) for, 284–285t
Molecular dynamics (MolD), 253, 256, 258–261. *See also* MolD simulation
  phases of, 259–260
Molecular evolutionary force, transposable elements as, 700–701
Molecular exclusion, chromatography, 35–36
Molecular graphics, 537–540
  programs, 95
Molecular imprinting, 529
Molecular interactions/docking, 253
Molecular ions, 50
Molecularly imprinted polymer (MIP), 529
Molecular mechanics (MM), 249, 250–252, 253–254
  computational portion of, 256
  energy minimization, 265–267t
  potential functions in, 255
Molecular model, construction from electron density map, 217. *See also* Molecular modeling entries

Molecular modeling (MM)
  classes in, 361
  molecular mechanical approach to, 252–264
  power of, 253
  programs, 264t
Molecular modeling database (MMDB), 607
Molecular modeling packages, computational application of, 263–264
Molecular orbitals (MOs), 187
Molecular phylogeny, elements of, 685–687
Molecular properties
  calculating, 253
  sequence/structure relatedness for, 277
Molecular recognition, in carbohydrate-lectin interaction, 312–320
Molecular strain energy, minimizing, 256
Molecular structure, determination of, 215–216
Molecular thermodynamics, 249–252
Molecular weight
  of biomacromolecules, 44–50
  empirical methods of determining, 48, 49t
  physicochemical methods of determining, 46–48
Molecules, configuration of, 9
Mono-ADP-ribosylation, 486
Monoantennary glycans, 171
Monoclonal antibodies
  as catalysts, 385
  complexed with protein antigens, 303–304
  preparing, 302–303
  as target molecules, 528
  therapeutic, 509t
Monodisperse sample, 46
Monoisotopic mass, 103
Monomer biomolecules, 28–29
  constituents of, 15–30
Monomers, 55
Monophyletic group, 686
Monosaccharide database, 27
Monosaccharide derivatives, 26–27
Monosaccharide residues, 148
Monosaccharides, 23, 24
  ^{13}CMR data for, 160–161, 207–208
  common, 25–26t
  as glycan constituents, 23–27
  hydrolysis to, 154
  linear codes for presenting, 150–151
  residue masses of, 158t
Monosaccharide sequences, determination of, 176–177
Monosaccharide unit connections, linear codes for, 151–153
Monte Carlo (MC) simulations, 261, 263
Moonlighting enzymes, 333t
Motifs, 604–605
MPDB Web site, 497
MRC Laboratory of Molecular Biology Web server, 610
mRNA codon:tRNA anticodon, decoding of, 479–481. *See also* Messenger RNA (mRNA)
mRNA expression, RNA interference with, 467–469
mRNA polyadenylation, 463
mRNA quantification, 583
mRNA transcripts, posttranscriptional processing/modification of, 469–471
MS–MS techniques, 158. *See also* Tandem mass spectrometry (MS-MS)
Mucin-type glycosylation, 173
Multidomain proteins, 703
  structures of, 135
Multienzyme complexes, 331–333
Multienzyme systems, 326
Multifunctional enzymes, 326, 333
Multipin peptide synthesis, 242
Multiple-base mismatches, 74
Multiple equilibria, 291–295
Multiple nucleotide/peptide/saccharide synthesis, 241–242
Multiple sequence alignment, 522–523
Multiple signal transduction pathways, calmodulin-initiated, 412
Multiple-site binding, 292
  equivalent sites in, 293–294
  nonequivalent sites in, 294–295
Multiple-site receptors, responses to binding plots, 295t
Multiplex PCR, 498. *See also* Polymerase chain reaction (PCR)
Multipurpose Internet Mail Extensions (MIME), 544
Multisubstrate enzyme reaction, rate equation for, 339–340
Multivalency, in lectin-carbohydrate recognition, 317–318
Muscle glycogen phosphorylase, allosteric effectors of, 381t
Mutagenesis
  nonsense-suppression, 643–647
  site-directed, 350, 499, 501–504
Mutarotation, 23, 24t
Mutation(s)
  methods used to scan genes for, 567t
  rate of occurrence of, 682, 683
  role in evolution, 680–682
  spontaneous, 74
Mutation data (MD) score, 519–520

N
NADH, 399. *See also* Nicotinamide entries
NADPH, 399
Nanoengineering, use of DNA in, 91
Nanospray technique, 102
Nascent polypeptide-associated complex (NAC), 489
National Center for Biotechnology Information (NCBI), 56, 547, 569. *See also* NCBI entries
National Library of Medicine (NLM), 547

Natural antisense transcript (NAT), 500
NCBI databases, 551. *See also* National Center for Biotechnology Information (NCBI)
NCBI Web-based tools, 583
NCBI Web site, 444
NEBcutter Web site, 562
Necrotic process, 419
Needleman–Wunsch algorithm, 536–537
Negative cooperativity, 295
Negative feedback controls, 378
Negative ion spectra, 102
Negatively supercoiled DNA, 79
Neighbor joining (NJ) method, 690
Neoglycoproteins, 177
  preparations of, 178–179t
NetEntrez, 551. *See also* Entrez entries
NetOGlyc server, 174
Network Protein Sequence Analysis (NPS@), 279, 617
NeuNAc interactions, 320
Neutral mutation rate, 683, 684
Newsgroups, 542–543
Newton–Raphson procedure, 257, 258
N-glycan synthesis, 670–671
N-glycosidic glycoproteins, 657
Nick-closing enzymes, 451
Nicked (form II) DNA, 78
Nicotinamide coenzymes, 362, 399
Nicotinamide nucleotide binding core, 281
90° pulse, 202, 203
Nitric oxide, 413–414
Nitric oxide synthase (NOS), 413
N-linked glycans, 172
N-linked glycoprotein synthesis, 439–440
N-linked oligosaccharide chains, 168–169
  complex type, 172
NMR spectra. *See also* Nuclear magnetic resonance entries
  multiplet structure of, 200–201
  recording of, 199
nnPredict Web site, 618
NOESY experiments, 204–205. *See also* Nuclear Overhauser effect and exchange spectroscopy (NOESY)
NOESY spectrum, 205
Nonbonded interactions, guidelines for, 256
Nonbonded repulsive interactions, 11
Nonbonding potentials, 255
Non-canonical base pairs, 86
Noncoding RNAs (ncRNA), 85
Noncompetitive inhibition, 342
Noncooperative transition model, 269–270
Noncovalent catalysis, 344–346
Noncovalent glycan immobilization, 676–677
Noncovalent interactions, 5–8
  DNA-protein, 305
  in Ig molecules, 301
Nonhistone chromosomal proteins, 140
Nonhomologous structures, catalytic site convergence in, 706–707
Nonlinear 2:2 function kinetics, 340–341
Nonlinear kinetics, 339–341
"Nonpolar-in, polar-out" rule, 132
Nonpolar interaction, in lectin-carbohydrate recognition, 316–317
Nonradiative quenching, 192
Nonreceptor protein tyrosine kinases (nrPTKs), signaling pathways operated by, 419
Nonrepetitive protein structures, 115–116
Nonsense suppression approach, 350
Nonsense suppression mutagenesis, 643–647
Nonspecific protein recognition, sugar-phosphate backbone and, 307
NPG-dependent glycosyl transferases, 437
NRSAS Web site, 403
N-terminal ladder sequencing, 101
N-terminal residue, determining, 97
Nuclear magnetic resonance (NMR), 159–161
  of glycans, 207–208
  of nucleic acids, 206
  of proteins, 203–205
  two-dimensional Fourier transform, 202–203
Nuclear magnetic resonance spectroscopy, 197–208, 353–356
  application of, 205
Nuclear magnetic resonance spectrometry, 184. *See also* NMR spectra
Nuclear magnetization, 198–199
Nuclear Overhauser effect (NOE), 202, 203
Nuclear Overhauser effect and exchange spectroscopy (NOESY), 203. *See also* NOESY entries
Nucleases
  cleavage pattern and specificity of, 426t
  essential functions of, 425
Nucleation-condensation protein folding model, 493
Nucleation parameter, 271
Nucleic acid abbreviations, 56
Nucleic acid analysis servers, 571
Nucleic acid constituents, 15–18
Nucleic Acid Database (NDB), 65, 216
Nucleic acid databases, 568–571
Nucleic acid fold, in structure prediction, 281–282
Nucleic acid interactions, complementarity in, 305–311
Nucleic acid melting, 275
Nucleic acid programmable protein array (NAPPA), 640
Nucleic acids, 2, 11, 15. *See also* Deoxyribonucleic acid (DNA); Protein-nucleic acid complexes; Ribonucleic acids (RNAs)
  absorption of, 188–189
  applications of, 90–91

  - base-pairing complementarity in, 281
  - as biocatalysts, 323
  - biomacromolecular structure of, 55–93
  - helical transition in, 274–275
  - hydrolytic degradation of, 424
  - IR spectral region of, 197
  - key structural features of, 63–66
  - melting temperature of, 90
  - NMR of, 206
  - nucleolysis of, 424–426
  - ORD and CD studies of, 213
  - purification of, 31, 39
  - self-splicing, 325
  - sequence analysis of, 57–63
  - spectral parameters for, 189t

Nucleic acid secondary databases, 573t

Nucleic acid sequence, evolution of, 697–699

Nucleic acid structures
  - energetics of, 89–90
  - representation of, 56–57

Nucleobase catalysis, 393–394

Nucleolytic ribozymes, 387, 388

Nucleophilic groups, in enzymes, 346t

Nucleophilic sites, for chemical modifications, 350

Nucleoproteins, quinternary structure of, 140–143

Nucleosides
  - pK_a values for bases in, 17t
  - spectral parameters for, 189t
  - in transfer RNA, 82

Nucleotide excision repair (NER), 459

Nucleotide library, 241

Nucleotide phosphates, 16

Nucleotides, 2
  - acid-base behavior of, 16
  - HPLC analyses of, 28t
  - interaction strength of, 491
  - as nucleic acid constituents, 15–18
  - pK_a values for bases in, 17t
  - shapes of, 18

Nucleotide sequences, 56–57
  - information content of, 563–564

O

O-Glycobase Web site, 662–663

O-glycosidic glycoproteins, 657

Okazaki fragments, 447

Oligodeoxynucleotide/DNA fragments, sequence analysis of, 42

Oligodeoxyribonucleotide/site-directed mutagenesis, 499

Oligomeric biomacromolecules, diagnostic tests for cooperativity of, 299t

Oligomeric ion channels, 402–403

Oligonucleotides
  - chemical synthesis of, 220
  - melting temperature for, 90
  - MS fragmentation of, 61
  - synthesis of, 236–237, 579–580

Oligopeptide synthesis, 232–236

Oligosaccharide chains
  - biosynthesis of, 436–437, 438–439
  - structure diversity of, 168–174

Oligosaccharide engineering, metabolic, 672–673

Oligosaccharides, 13, 149, 657–659
  - complex, 170
  - high mannose, 169–170
  - hybrid type, 170
  - NMR in analyses of, 207
  - “stop point” fragments of, 177
  - synthesis of, 237–241

Oligosaccharyltransferase (OST), 439

O-linked glycans, 172–173

O-linked glycoprotein synthesis, 440

O-linked oligosaccharide chain, 169

Oncogenic viral gene products (vRas), 406

Oncology antibodies, 510

One-dimensional random walk, 271–273

1D presentations, 538

180° pulse, 202

Online bibliographies, 578

Online phylogenetic analysis, 697

Online protein structure prediction, 626–628

Ontology, gene, 553–554

Open β-α-β proteins, 135

Open reading frame (ORF), 563

Open-source software, 537

Operating System (OS), 535–536

Operational taxonomic units (OUTs), 685–686

Optical activity
  - detecting, 208
  - side chain, 212

Optical isomerism, 9

Optical rotatory dispersion (ORD), 208–209. *See also* ORD entries
  - empirical applications of, 212–214
  - main applications of, 213

ORD curves, 209, 210

ORD spectra, 208, 209–212

Organosilane compounds, in surface preparation, 526

Orthogonal β-sheet proteins, 135

Orthogonality, protecting, 221

Orthogonal methods, 633–634

Orthogonal protecting groups, selecting, 237

Oscillatory effect, 377

Osmotic pressure measurement, in determining molecular weight, 46–47

Oxidases, 323

Oxidation reactions, 366

Oxidoreductases, 323, 361–366

Oxygenases, 366

Oxygenation reactions, 366

P

Pair-wise alignment, 517, 689

Palindromes, 74, 458

PANAL Web site, 605

Papain, 426
  - catalysis, 428

Parallel analysis, of monosaccharide sequences, 177

Parallel β-sheet proteins, 135

Parallel β-sheets, 114

Parallel synthesis, 241–242

Parallel tetraplex structure, 77

Paralogues, 517

Parsimony searches, 691
Partition chromatography, 37
Partition coefficient, 360
Partition function, for the zipper model, 270
Pasteur effect, 377–378
Pathway analysis, 581
Patterson map, 217
PCR amplification, 498, 578. *See also* Polymerase chain reaction (PCR)
Pdb21inucs Web site, 664
PDBSum Web site, 330, 607–608. *See also* Protein Data Bank (PDB)
Pectins, 165
Pentaantennary glycans, 171
PepConfDB, 108
Peptide bonds, 22, 187–188. *See also* Polypeptide entries
  properties of, 109
Peptide chains, assigning secondary structures along, 278
Peptide cleavage, 235
Peptide cleavage sites, identification of, 634
Peptide ladder sequencing, 101
PeptideMass program, 597
Peptides
  cleavage, separation, and analysis of, 97
  cyclic, 130
  determination of partial amino acid composition of, 633
Peptide segments, condensation of, 233
Peptide sequence, determination of, 99
Peptide synthesis
  photolithographic, 242
  protection and deprotection specific to, 223–225
  solid phase, 225
Peptidoglycans, 167
Periodate oxidation, 155
Periodic box, 249–250
Perturbation difference absorption spectroscopy, 189–190
PEST sequences, 429–430
  searches for, 616
Pfam protein domain families database, 605
Phage display antibody library, 639–640
Phage display library, 507–508
Phagemid, 507
pH effect, in enzymatic reactions, 343–344
Phenolic hydroxylation, 366
Phenotype, 560
Phenylalanine, characteristics of, 21t
Phosphite triester synthesis, 237
Phosphodiesteric/peptidic/glycosidic bond formation (coupling reaction), 225
Phospholipase A_2 (PLA_2), 408
Phospholipase C (PLC). *See also* PLCγ
  β forms of, 409
  in signal transduction, 408–410
Phospholipase D (PLD), 408
Phosphopeptide sequencing, 635
Phosphoprotein-ubiquitin ligase complexes (PULCs), 432
Phosphorolysis, of glycans, 422–424
Phosphorus magnetic resonance spectroscopy, 356
Phosphorylation, in signaling, 414–417
Phosphorylation/dephosphorylation regulatory mechanism, 375
Phosphorylative allosteric transition, 381
Phosphotriester synthesis, 236
Phosphotyrosine binding (PTB) domain, 418
Photodeprotection, 242
Photolithographic spatially addressable multiple peptide synthesis, 243
Photolithographic techniques, 242
Photolithography, in oligonucleotide synthesis, 579–580
pH-Rate profiles, 367
Phylogenetic analysis, 685–686
  of biosequences, 687–693
  choice of methods for, 689–690
  online, 697
  of SINE, 699
  software, 693–696
Phylogenetic databases, 697t
Phylogenetic inference, sequence analyses in, 693–697
Phylogenetic Inference Package (PHYLIP), 693–696
Phylogenetic method
  character-based, 691–692
  distance-based, 690–691
Phylogenetic results, assessment of, 692–693
Phylogenetics, 686–687
Phylogenetic trees
  constructing, 692
  representations of, 696
Physical gene mapping, 561–562
Physicochemical probes, 647
Pimer3 Web site, 496
Ping pong mechanism, 339
PIR-International Protein Sequence Database (PSD), 599
PIR Protein Sequence Database, 599
pK_a values
  for bases in nucleosides and nucleotides, 16, 17t
  estimation of, 343–344
Planar aromatic chemicals, binding to DNA, 309–310
Plant lectins, 319
PlasMapper Web site, 496
Platelet derived growth factor (PDGF), 417
PLCγ, 417–418. *See also* Phospholipase C (PLC)
P-loop, 127
PMR spectra, 203–204. *See also* Proton magnetic resonance (PMR)
  of nucleic acids, 206
  of oligo-/polysaccharides, 207
P—O ester bonds, 18
Point accepted mutation (PAM), 519–520
Pol I polymerase, 449–450
Pol II polymerase, 450
Pol III polymerase, 450
Polak–Ribiere algorithm, 257
Polarimetric measurement, 157

Polyacrylamide gel electrophoresis (PAGE), 41. *See also* High-resolution polyacrylamide gel electrophoresis (hrPAGE); Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE); SWISS-2DPAGE Web site; WORLD-2DPAGE index
  two-dimensional, 43, 44
Polyacrylamide gels, 42
Poly-ADP-ribosylation, 486
Polyamino acids, 3
Poly(dimethylacrylamide), cross-linked, 230
Polydisperse sample, 46
Polymerase action, characteristics of, 448–449
Polymerase chain reaction (PCR), 236, 496–499. *See also* PCR amplification
  variations on, 498–499
Polymerization, initiation of, 462–463
Polymer support, solid-phase, 230–231
Polynucleotides, 1–3
  synthesis of, 236–237
Polypeptide chain ligation, biochemical, 245
Polypeptide chains, temperature-dependent transition of, 273–274
Polypeptides, 3. *See also* Peptide entries
  coil-helix transition in, 273–274
  synthesis of, 232–236
Polypurine-polypyrimidine tracts, 76–77
Polysaccharide chains
  biosynthesis of, 436–437
  secondary and tertiary structures of, 161–163
Polysaccharides, 3, 13, 147–182. *See also* Glycans
  biological functions of, 147
  classes of, 148
  NMR in analyses of, 207
  primary structure of, 153–161
  structures of, 163–166
  synthesis of, 237–241
Porin, 128
Positional isomerism, 8
Position sensitive iterated-basic linear alignment sequence tool (PSI-BLAST), 523. *See also* Basic Local Alignment Search Tool (BLAST) program; PSI-BLAST searches
Position weight matrix (PWM), 577
Positive cooperativity, 295
Positively supercoiled DNA, 79
Post-replicational modification (PTM), 456–458
  sites of, 614–616
Post-translational modification analysis, by mass spectrometry, 634–635
Post-translational protein modifications, 484–488
  common, 485t
Post-translational translocation, 489, 490
Potential energy, 251
PPSearch Web site, 604
Practical Extraction and Report Language (Perl), 536
Precipitation, in biomacromolecule purification, 34
Precursor mRNA (pre-mRNA), 85
PRED-GPCR Web site, 403
PredictProtein Web site, 613
Pre-existing DNA, as a template, 445
Prehenylation, 484–485
Primary databases, 549–550
Primary macromolecular structure, 55, 180t
Primary metabolism, 398
Primary monosaccharide specificity, in carbohydrate recognition, 318–319
Primary structures
  protein, 95–108
  tRNA, 82–83
Primary sequence databases, 598–602
Primary shift, 200
Primase, 454
Primer design, automated, 578–579
Primer design tool, 584
Primer dimers, 497–498
Primer extension preamplification (PEP), 498–499
Primers, for polymerase chain reaction, 496–497
Primosome, 453–454
Principal component analysis (PCA), 532
PRINTS fingerprint database, 604–605
PROBEmer Web site, 496
Probes, microarray, 529–530
Pro-caspases, 422
PROCAT Web site, 330
Processive enzymes, 436
Prodock tool, 285
Product ion spectra, interpreting, 106–108
Proenzymes, 374
Profiles sequence similarity search, 523
Programmed cell death (PCD), 419. *See also* Apoptosis
Programming languages, 536
Prokaryotes, daughter strand identification in, 457
Prokaryotic genes, 559
Prokaryotic genome organization, 560
Prokaryotic polymerases, properties of, 449–450
Prokaryotic proteasomes, 432
Prokaryotic ribosomes, 141t
Prokaryotic RNA transcription, 461–463
Prokaryotic transcription
  control circuits for regulation of, 465t
  regulation of, 466
Prolactin receptors, 419
Proliferating cell nuclear antigen (PCNA), 450
Proline, characteristics of, 19t
Promoters, detecting, 577
ProNIT Web site, 306
Propeller twist, 65
Proproteins, 490
PROPSEARCH Web site, 596
PROSITE database, 615
PROSITE Web site, 604
PROSPECT-PSPP Web site, 627
ProtDist program, 694

Proteases
  active site of, 426
  mechanistic groups of, 428t
Protection/deprotection
  of common functional groups, 221–223
  specific to peptide synthesis, 223–225
Protection reactions, 221
  types of, 221–223
Protection schemes, in peptide synthesis, 233–235
Protein, post-translational modifications of, 484–488, 595. *See also* Proteins
Protein absorption, 187–188
Protein-based approach, to predicting functional sites, 280
Protein binding, folds and, 126–127
Protein biosynthesis, 472–490
Protein chain ligation, biochemical, 245
Protein chemistry, mass spectrometry in, 101–103
Protein complexity evolution, 702–703
  domain duplication and, 703
Protein constituents, 18–23
Protein Data Bank (PDB), 95, 111, 216, 605–607, 662. *See also* PDBSum Web site
  file format for, 540, 541
Protein degradation pathway, 427–433
Protein detection schemes, 43
Protein domains, 702
  classification of, 136
  structures of, 134
Protein engineering, 501–505
Protein expression analysis, approaches to, 595
Protein fluorescent spectra, rules for interpreting, 191
Protein folding, 491
  DNA-induced, 305
  mechanisms of, 493–494
Protein folding pathway, *in vitro*, 492–494
Protein folds
  changes in, 704–706
  tertiary structures and, 121–126
Protein functional group properties, influences on, 350
Protein functional sites, 280–281
Protein function evolution, 706–708
Protein glycosylation, 634
Protein helical segments, optical activity of, 211
Protein hormones, 12
Protein identification, by mass spectrometry, 632
Protein identity
  composition- and properties-based, 595–596
  sequence-based, 596–597
Protein IR bands, characteristics of, 196t
Protein kinase A (PKA), 414–415
Protein kinases, 414–416
Protein kinases C (PKCs), 415–416
Protein-ligand interactions, 300
Protein microarrays (chips), 638–640
  preparation of, 528
Protein Modification Screening Tool (ProMoST), 634
Protein molecules, architecture of, 94–95
Protein-nucleic acid complexes, prominent features in, 306
Protein phosphorylation, studying, 634–635
Protein profiling, quantitative, 639
Protein–protein interactions, 628–629
  analysis by two-hybrid assay, 636–637
Protein recognition
  major- and minor-groove bases available for, 307
  nonspecific, 307
Proteins, 3, 11. *See also* Enzymatic entries; Enzyme entries; Peptide entries; Polypeptide entries
  allosterism and cooperativity in, 295–299
  as biocatalysts, 323
  cell surface, 136
  charge of, 22
  chromosomal, 140
  circular (cyclic), 129–130, 131t
  classification of, 94
  conformational map for, 109
  domain structures in, 119–121
  evolutionary rates of, 683–684
  fibrous, 128–129
  fold recognition and threading in, 623
  functions of, 12–13, 127
  glycosylation of, 486–487, 665–666
  interaction with RNA, 310–311
  irregular and small, 136–137
  isoelectric pH of, 52–53
  membrane, 128, 136
  multienzyme, 331–333
  NMR of, 203–205
  ribosomal, 141, 142
  versatility in, 12
Protein secondary structures, 110–118, 209–212. *See also* Protein structure entries
  predicting, 617–618
  structure prediction from, 277–279
Protein sequence analysis, tandem mass spectrometry in, 103–108
Protein Sequence Analysis (PSA) server, 618
Protein sequence secondary databases, 603t
Protein splicing, 245, 487–488
Protein stability, enhancing, 503–504
Protein structural flexibility, 285
Protein structure(s). *See also* Structural entries; Structure entries
  classification of, 133–137
  databases of, 609t
  evolution of, 701–703, 704–706
  multidomain, 135–136
  nonrepetitive, 115–116
  primary, 95–108
  quaternary (subunit), 137–139
  quinternary, 140–143

- random coil conformation of, 212
- representation of, 94–95
- similarity/overlap of, 620–622
- supersecondary, 117–118
- tertiary, 118–133

Protein structure analysis, using bioinformatics, 616–629
Protein structure databases, 609t
Protein structure modeling, 3D, 618–619
Protein structure prediction
- *ab initio*, 618–619, 624–625
- from amino acid sequences, 616–617
- approaches to, 279
- on-line, 626–628

Protein topology, representation of, 130. *See also* Protein structure(s)
Protein topology cartoons (TOPS), 130
Protein translation, 472–474
- eukaryotic initiation factors for, 476t
- mechanics of, 474
- processes in, 475–479

Protein translocation, 488–490
Protein trans-splicing, by employing split inteins, 245, 246
Protein tyrosine kinases (PTKs), classes of, 417
Protein tyrosine phosphatases (PTPs), 416
Proteoglycans, 167
Proteoglycomics, glycoenzyme-based, 670–672
Proteolysis, 426–427
Proteolytic cleavage, 487
Proteome analysis/annotation, 610–613
- by mass spectrometry, 631–634

Proteome databases, 596t
- integrated, 613–614

Proteome expression/function, investigation of, 629–647
Proteome informatics
- proteomic servers, 610–616
- sequence databases and servers, 598–605
- structure databases and servers, 605–610

Proteomic research, 595–596
Proteomics, 515, 594–654
- investigation of proteome expression/function, 629–647
- metabolome and, 647–649
- protein structure analysis using bioinformatics, 616–629
- proteome features and properties, 594–597

Proteomic servers, 610–616
Proteomic Web servers, 611t
ProTherm Web site, 597
PROTINFO Web site, 625, 628
Protonated peptides, 104
Proton magnetic resonance (PMR), 198, 199. *See also* PMR spectra
- high resolution, 206
- spectroscopy, 159

Proto-oncogene products, 405
ProtPars program, 695
ProtScale program, 597, 617, 619
P selectin, 314
Pseudoknot constraints, 86
Pseudoknot structural unit, 84
PSI-BLAST searches, 619. *See also* Position sensitive iterated-basic linear alignment sequence tool (PSI-BLAST)
Public databases, 549
PubMed, 546
Pulsed field gel electrophoresis (PFGE), 42
Purine bases, fluorescence of, 192
Purines, 15, 16
- nucleic acid absorption and, 188

Purine stacks, cross-strand, 87
Purity, of biomacromolecules, 44, 45t. *See also* Biomacromolecule purification
Push-pull catalysis, 327
P-value, 524
Pyranose rings, six-member, 24, 161
Pyridoxal enzyme reaction, spectral changes accompanying, 355
Pyridoxal enzymes, 368–372
Pyrimidine bases, fluorescence of, 192
Pyrimidine dimers, enzymatic photoreactivation of, 459
Pyrimidines, 15, 16
- nucleic acid absorption and, 188

Pyrimidine tract-binding (PTB) proteins, 142, 143
Pyruvate dehydrogenase complex (PDC), 331–332

Q

Quantitative structure-activity relationship (QSAR), 357–358, 360
Quantum mechanics (QM), 250
Quantum yield, 190
Quasi-catalytic biochemical reactions, 323
Quasi-equilibrium assumption, 334
Quasi-equilibrium rate equation, writing, 338
Quasi-equilibrium treatment, of random reactions, 338
Quaternary macromolecular structure, 55, 180t
Quaternary (subunit) protein structures, 137–139
Quenched dynamics, 260
Quenching, 191, 192
Quinternary structures
- macromolecular, 55, 180t
- of nucleoproteins, 140–143

R

Radiation
- absorption of, 183–184
- emission of, 185

Raf protein phosphorylating enzyme, 406
Raft-associated GPI proteins, 173
Ramachandran plot, 108–110
Raman spectrum, 193
Random coil, 273
Random-coil conformations, CD spectra of, 212
Random reactions, quasi-equilibrium treatment of, 338

Random structures, modeling of, 271–273
Rapid-equilibrium assumption, 334
Rapid synthetic methods, 241
RasMol, 218, 264
Ras proteins, 405–406
Ras signaling pathway, 418–419
Raw score, 524
Rayleigh scattering, 47
Reaction mechanisms, information sources for, 349
Reaction path shifting, 505
Reaction quotient, 289
Reaction specificity, 328
Reactive group, in activity-based probes, 642
Reactive intermediate sequestration, by multienzyme complexes, 332
ReadSeq Web site, 538
Reagent array analysis method (RAAM), 177
Real time PCR, 499. *See also* Polymerase chain reaction (PCR)
Real-time rotation/clipping, 538
REBASE Web site, 458
RecA protein, 461
Receptor Database, 400
Receptor protein tyrosine kinases, signal pathways operated by, 417–419
Receptors, 400–403
  tyrosine kinase-containing, 417–418
Recoding, modes of, 481–483
Recognition helices, 306–307
Recombinant DNA technology, 494–499
Recombinant glycoproteins, 673–674
Recombinase, 461
Redox coenzymes, 361–362
Redox reactions, categories of, 362–366
Reductive carboxymethylation, 366
Reflection, radiation scattering via, 183
Refraction, radiation scattering via, 183
RefSeq Web site, 614
Regular structure transitions, two-state models of, 268–271
Regulatory enzymes, 326
  covalent interconversion of, 375
Regulons, 308
Relaxation times, 201–202
Relaxed DNA, 78
Released glycans, labeling, 175
Released oligosaccharides, purification of, 176
Relibase, 616
REM-TrEMBL, 600
Repetitive DNA masking, 575–576
Replication fork, 446–447
Replication termination protein, 448
Replicons, multiple, 448
Repressors, 466
Research Collaboratory for Structural Bioinformatics (RCSB), 540
Residue-by-residue alignment, 517–518
Residue mass values, 105–106
Residues per turn, 65
Residue types, structural states adopted by, 278
Resins
  linkers used in, 231t
  peptide cleavage from, 235
Resonance, 22–23
Resonance intensities (peak heights), 199
Resonance Raman spectroscopy, 353
Resources, Internet, 540–546, 546–548
Restriction endonucleases, 457
Restriction enzymes, 57, 458
Restriction maps, 562–563
Retaining glycosidases, catalysis by, 424
Retroviruses, 455, 482
Reverse phase chromatography, 36
Reverse transcription, 455–456
Reverse transcription polymerase chain reaction (RT-PCR), 498. *See also* Polymerase chain reaction (PCR)
Reversible inhibition patterns, 342t
Reversible inhibitor, 341
Reversible protein phosphorylation, 634
R-factor, 218
Ribonuclease H activity, 455
Ribonuclease P (RNase P) RNA, 391–392
Ribonucleic acids (RNAs), 2, 57. *See also* Messenger RNA (mRNA); Ribosomal RNA (rRNA); RNA entries; Transfer RNA (tRNA)
  annealing and melting, 275
  cellular functions of, 81–82
  structures of, 81–85, 86–89
Ribonucleoproteins, 387
Ribonucleotides, 15
Ribose zippers, 89
Ribosomal RNA (rRNA), 81, 83–84, 141, 392. *See also* rRNA folding
  functional roles of, 84
Ribosomes, 141–142
  stalled, 483
Riboswitch, 466
Ribozyme catalysis, strategies for, 392–394
Ribozymes, 386–394
  characteristics of, 387t
  description of, 388–392
  engineering, 500–501
Rms displacement, 272. *See also* Root-mean-square entries
Rms radius, 272
RNA-directed DNA polymerase activity, 455. *See also* Ribonucleic acids (RNAs)
RNA editing, 471
RNA engineering, 500–501
RNA folding, 86, 491–492
RNA-induced silencing complexes (RISCs), 469
RNA interference (RNAi), 467–469
RNA intermediates, role of, 701

RNA polymerases
  binding to DNA template, 462
  in eukaryotic cells, 463–464
RNAP promoters, 464
RNA-protein interaction, 310
RNA-recognition motif (RRM), 141
RNase P cleavage reaction, 392
RNA sequence databases, 572t
RNA structures, approaches to predicting, 282
RNA transcript, polymerization of, 463
RNA transcription, 461–463
  eukaryotic, 463–464
  prokaryotic, 461–463
  regulation of, 466–469
Rolling-circle replication, 448
Root-mean-square deviation (RMSD), measuring, 621–622
Root-mean-square (rms, RMS) distance, 622
Root-mean-square end-to-end distance, 272
Root-mean-square gradient, 254
Rossmann nucleotide-binding fold, 127
Rotation per residue, 65
Rotation symmetry, 137–139
rRNA folding, 492. *See also* Ribosomal RNA (rRNA)
Ryanodine receptors (RyRs), 410

S
Saccharide-binding sites, 315
Saccharide biosynthesis, 436–442
Saccharides, enzymatic solid-phase synthesis of, 239–241
Salt bridge, 305
Salting-in/-out ions, 36
SANDOCK tool, 285
SA-Search Web site, 608
ScanProsite tool, 604
Scatchard equation, 293
Scatchard plot, 295
Scattered intensity, 214
Scattered wave, intensity of, 215
Scatter plots, 532
Schiff base, 369
Scoring function, in molecular docking, 283
Screw symmetry, 139
Search algorithm, in molecular docking, 283
SearchFields, 606, 607
SearchLite, 606, 607
Secondary databases, 551
Secondary macromolecular structure, 55, 180t. *See also* Secondary structure entries
Secondary metabolism, 398
Secondary sequence databases, 598, 602–605
Secondary shift, 200
Secondary structure characteristics, correlation to sequence preference, 279. *See also* Secondary macromolecular structure
Secondary structure predictions, 617–618
  approaches to, 276–277
Secondary structures. *See also* Protein secondary structures
  base-pairing and, 281–282
  DNA, 63–77
  globular proteins, 116–117
  glucan, 162
  peptide chains, 278
  polysaccharides, 278
  tRNA, 82–83
Second messengers, 400–402, 403
Sedimentation, in determining molecular weight, 47–48
Sedimentation equilibrium, 48
Sedimentation velocity, 47
Segment-oriented prediction, 617
Selective antibody catalysis, 383
SELEX_DB Web site, 587. *See also* Systematic evolution of ligands by exponential enrichment (SELEX)
SELEX protocol, 585–586
Self-consistent field (SCF) methods, 250
Self-splicing introns, 389
Self-splicing nucleic acids, 325
Semi-conservative DNA replication, 446
Semi-empirical methods, 250
SEQANALREF Web site, 564
Seqboot program, 696
Sequence alignment, 517
  evolutionary basis of, 517–518
  methods of, 687–688
Sequence analyses
  methods of, 99–101
  of nucleic acids, 13, 57–63
  in phylogenetic inference, 693–697
  of polysaccharides, 153–161
  of proteins, 13
  by tandem mass spectrometry, 101–108
Sequence comparison
  domains in, 120
  dot matrix representation in, 518–519
Sequence data, in phylogenetic analysis, 687–690
Sequence databases/servers, 550, 598–605
Sequence motifs/patterns, 119t
Sequence Retrieval System (SRS), 547–548, 552, 601. *See also* SRS Web site
Sequences, linear codes for presenting, 150–153
Sequence search/alignments, statistical significance of, 524–525
Sequence similarity/alignment, 517, 619–620, 703
Sequence-specific RNA-binding proteins, 311
Sequencing strategies, 516
Sequential analysis, of monosaccharide sequences, 176–177
Sequential mechanism, 339
Sequential model, of homotropic interactions, 297–298
Serial analysis of gene expression (SAGE) technique, 584
Serine, characteristics of, 19t
Serine proteases, 366, 368, 426
Serine/threonine kinases, 414
Serine-threonine phosphatases, 416–417
Seven-transmembrane (7TM) segment receptors, 400–402

Shine-Dalgarno sequence, 475
Short interspersed nuclear element (SINE) sequences, 698–699
Short tandem repeat polymorphisms (STRPs), 565
Sialoadhesin, 314
Side chains, hydrophobicity versus hydrophilicity of, 279
Signaling, phosphorylation and dephosphorylation in, 414–417
Signaling pathways
  operated by nonreceptor protein tyrosine kinases, 419
  operated by receptor protein tyrosine kinases, 417–419
Signal transduction, 398–400
  adenylyl cyclase in, 406–408
  calcium signaling, 410–414
  elements of, 400–406
  phospholipase C in, 408–410
Silk fibroin, 129
Similarity scoring, 519–520
Similarity search approach, 522–524
  to functional site residue mapping, 281
Simulated emission, 185
Simulations
  molecular dynamics, 260, 267
  Monte Carlo, 261
Single-base mismatches, 73–74
Single nucleotide polymorphisms (SNPs), 567–568, 659
Single-site binding, 291
Single-stranded DNA (ssDNA), 453–454
Single-stranded genomes, 558
Single-transmembrane segment catalytic receptors, 400
siRISC, 469
siRNAdb Web site, 500
Site-directed mutagenesis, 245, 350, 499
  effects of, 502–503t
  enzyme engineering by, 501–504
  studies of, 353t, 354t
Site-specific chemical modifications, 633
Skeletal models, computer, 537–540
SLAM Web site, 575
Slater type orbitals (STO), 250
SLIDE tool, 285
Slipped, mispaired DNA (smp-DNA), 74
Small interfering RNA (siRNA), 85, 467–469
Small nonmessenger RNAs (snmRNA), 85
Small nuclear ribonucleoprotein (snRNP), 142
Small nuclear RNAs (snRNAs), 85, 142
Small nucleolar RNAs (snoRNAs), 85
Smart drugs, 508–511
SMART library, 605
SMILES, structural graphics using, 538
Smith–Waterman algorithm, 536–537
SNP databases, 568. *See also* Single nucleotide polymorphisms (SNPs)
Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE), 41, 42, 629–630. *See also* Polyacrylamide gel electrophoresis (PAGE)
Software. *See also* Freeware
  molecular modeling, 264
  open-source, 537
  phylogenetic analysis, 693–696
Solid phase synthesis (SPS), 675
  of biopolymers, 226–230
  chemical database for, 230
  glycoside, 238–241
  oligo- and polynucleotide, 236–237
  oligo- and polypeptide, 232–236
  oligo- and polysaccharide, 237–241
  peptide, 234
  phosphoramidite nucleotide, 238
  practice and application of, 232–241
  strategy for, 225–231
Solid-phase polymer support, 230–231
Soluble RNA (sRNA), 81, 82–83
Solution(s)
  biomacromolecules in, 289–291
  radiation scattering in, 183
Solvation, 625–626
Solvent entropy, 262, 263
Solvents, in biomacromolecule purification, 34
SOS mutagenesis system, 461
SOS translesion DNA synthesis, 461
Southern hybridization, 699
Spacer molecule, 38–39
Specificity constant, 335
Spectral Database System (SDBS), 28, 185
Spectroscopic methods, 183–185
  in biomacromolecular characterization, 186t
  choice of, 185
Spectroscopic probes, 647
  active site, 352–356
Spectroscopic techniques, 28
Spectroscopy
  biochemical, 183–185
  circular dichroism, 208–209
  fluorescence, 190–192
  infrared, 193
  nuclear magnetic resonance, 197–208
  ultraviolet and visible absorption, 185–190
  X-ray diffraction, 214–219
Spin-lattice relaxation, 201
Spin–spin interactions
  couplings, 200–201
  J splittings, 203
Spin–spin relaxation process, 201–202
Spliceosome, 142, 389–390
Splicing, 142–143
  protein, 487–488
Split-and-combine (split-and-pool) combinatorial synthesis, 242

Split-and-pool combinatorial nucleotide synthesis, 244
Split inteins, protein trans-splicing using, 245
Spontaneous emission, 185
Spray-ionization techniques, 102
SP-TrEMBL, 600
Src homology region 2 (SH2) domains, 418
SR proteins, 142–143
SRS Web site, 523. *See also* Sequence Retrieval System (SRS)
SSThread Web site, 618
Stable isotope tagging, 669
Stalled ribosomes, rescue system for, 483
Standard enthalpy change, 289
Standard entropy change, 289
Starches
  crystalline properties of, 164
  polysaccharide types in, 163–164
Stark degradation, 99
Statistical mechanics, 264
Statistical methods, structure prediction from sequence by, 276–282
Statistical thermodynamics, 264–273
STCDB Web site, 400
Steady-state assumption, 334
Steady-state kinetic treatment, of enzyme catalysis, 336–337
Steady-state rate equation, 336–337
Steepest descent method, 257
Stepwise chain assembly peptide synthesis, 233
Stereochemical complementarity, 118, 119
Stereoisomers, 9
Stereospecificity, 329
Steric exclusion, 319
Steric repulsion, 8
STING Millennium Suite (SMS), 608
Stokes radius, 51
Stokes shifts, 192
Stop codons, 444, 643
  redefinition of, 483
Strain factors, conformational, 10–11
Structural Classification of Proteins (SCOP) database, 133, 609–610
Structural divergence, measure of, 623
Structural enzyme reaction mechanism studies, 344
Structural genomics, 561
Structural glycans, 13
Structural graphics, levels of, 538
Structural isomerism, 8
Structural model, least square refinement of, 217
Structural transitions
  coil-helix, 273–274
  examples of, 273–276
  helical, 274–275
  topological, 276
  two-state models of, 268–271
Structural visualization, 252
Structure-activity relationship, of a chemical reaction, 357–360
Structure databases, 605–610
  substructures and structure classification in, 608–610
Structure prediction, from sequence, 276–282
Structure retrieval/generation, 252
Structures, superposing, 621
Structure superposition/alignment, 253
S-type mechanism (salt dependent), 75
Subsite multivalency, in lectin-carbohydrate recognition, 317–318
Substituent constants, 358, 359t
Substrate channeling, by multienzyme complexes, 332
Substrate saturation curve, 335
Substrate specificity, conversion of, 505
Substrates/products, inhibition/activation by, 378
Subtiloligase, 505
Subunit multivalency, in lectin-carbohydrate recognition, 317–318
Sugar-phosphate backbone, in nonspecific protein recognition, 307
Sugar puckering, 18, 68, 69
Suicide substrate probes, 642
Supercoiled DNA, 77–80
  binding of intercalation agent to, 309–310
Supercoiling, 276
  role in gene expression, 79–80
  superhelical density and energetics of, 80
Superfamilies, 702–703
Superfolds, 124–126
  characteristics of, 125–126t
Supersecondary structures, in proteins, 117–118
Suppression, 482
Surface preparation, for microarrays, 525–527
Svedberg equation, 47
SWEET2 modeling service, 216, 662, 663
SweetDB Web site, 173–174, 662
SWISS-2DPAGE Web site, 610
Swiss Institute for Experimental Cancer Research (ISREC), 615
Swiss Institute of Bioinformatics (SIB), 599, 604
SWISS-MODEL repository, 610, 624, 627–628
SWISS-PROT database, 95, 596, 599–600, 610
Switch II region, 127
Switch regions, 406
Symmetry, structural, 137–139
Symmetry model, 378–379
  of homotropic interactions, 296–297
Syn-anti conformation, 18
Synthases, 324
Synthesis, combinatorial, 241–244
Synthetase complexes, studies of, 311
Synthetases, 324
Synthetic strategy
  conventional approach to, 220–225

  solid phase approach to, 225–231
Systematic evolution of ligands by exponential enrichment (SELEX), 91, 584–587

T
Taft equation, 359
Tagged image file format (TIFF) files, 531
Tagged proteins, 640
Tags, in activity-based probes, 642–643
Tandem-in-space mass spectrometers, 104
Tandem-in-time mass spectrometers, 104
Tandem mass spectrometry (MS-MS), 61, 632–633. *See also* Mass spectrometry (MS); MS–MS techniques
  in protein sequence analysis, 103–108
  sequence analysis by, 101–108
Taq polymerase, 497
Target-probe hybridization interactions, 580
TargetP Web site, 613
Target reaction transition state, mimicry of shape and electrostatic properties of, 383–384
TATA-binding protein (TBP), 464
TATA box, 464, 467
Tautomerism, 15–16
  resonance and, 23
T-DNA, 69
Telnet, 543
Telomerase, 454–455
Telomeres, 448, 454
Temperature effect, in enzymatic reactions, 343
Template approach, to functional site residue mapping, 281
TentaGel hydrophilic polymer, 230–231
Terminal determination, 97–98
Terminal loop RNA motifs, 87
Terminal oligosaccharide sequences, 658t
Termination signals, detecting, 577
Tertiary RNA motifs, 86
Tertiary structures
  cellulose chains, 166
  classification of, 122t
  DNA, 77–80
  macromolecular, 55, 180t
  polysaccharide chains, 161–163
  protein, 118–133
  protein folds and, 121–126
  tRNA, 83
Tetraantennary glycans, 171
Tetraloop-helix interactions, 89
Tetraloops, 87
Tetraplex DNA, 77
Text compression algorithms, 548
Therapeutic antibodies, 508–511
  modes of action of, 510
  monoclonal, 509t
Thermal (radiationless) emission, 185
Thermodynamic properties, computing, 263
Thermodynamics
  molecular, 249–252
  statistical, 264–273
Thiol-disulfide isomerization, 487
Thiol protecting groups, 227t
30S domain closure, 481
30S ribosomal subunit, 480
Threading, 623
3_{10} helix, 113
3D biomacromolecular structures, predicting, 282
3D graphics, 540, 542
Three-dimensional structure modeling, 618–619
3D-PSSM server, 626
Three-point attachment model, 329
Threonine, 18
  characteristics of, 19t
Through-bond interactions, 200
Through space interaction, 202
TIGR Gene Indices Web site, 575
Time-of-flight (TOF) mass analyzer, 632
T lymphocyte receptor (TCR), 419
TMAP Web site, 597
TMpred Web site, 597
Tn antigens, 173
Top-down gene mapping, 562
Topoisomerases, 451–453
Topoisomer free energy, 276
Topoisomers, DNA, 77–80
Topological switch points, 127
Toroidal coils, 80
Torsion angles, 4–5
Tosylation, 221
Total constructive interference, 215
Total genome amplification method, 498–499
Trajectory calculation, 260
TRANFAC Web site, 463
Transcription, steps in, 461–462
Transcription activation domain (AD), 636
Transcriptional regulatory proteins, 308
Transcriptome analysis, 583
Transducers, 403–406
Transferases, 323–324
Transfer-messenger RNA (tmRNA), 483
Transfer RNA (tRNA), 81, 82–83, 311, 472. *See also* tRNA entries
  aminoacylation of, 372
  folding of, 491–492
  orthogonality of, 643
  primary and secondary structure of, 82–83
  tertiary structure of, 83
Transfer RNA transcripts, posttranscriptional processing/modification of, 471–472
Transhydrogenation, 363–364
Transient-response method, 799
trans interactions, 301
Transition dipole moment, 193, 195
Transitions of regular structures, two-state models of, 268–271
Transition state, suitable microenvironment for, 384–385
Transition state ensemble (TSE) structures, 493

Transition state enzyme mechanism Studies, 356–357
Transition state inhibitors approach, 357
Translated EMBL (TrEMBL), 600–601. *See also* EMBL nucleotide sequence database
Translation
  over coding gap, 483
  protein, 472–490
Translation initiation site, detecting, 577
Translocation, protein, 488–490
Translocation pathways, 489
Translocons, 489
Transmembrane protein tyrosine phosphatases, 416
Transmission Control Protocol (TCP), 540
Transpeptidation, 477
Transport proteins, 124
trans-translation, 483
TREEALIGN Web site, 689
Tree of life, 698
TreeView program, 695
Triantennary glycans, 171
Trifluoromethane sulfonate hydrolysis, 175
4,5′,8-Trimethylpsoralen, binding to DNA, 310
Trinucleotides, 241
Triplex DNA, 73, 75–77
tRNA anticodon, 479. *See also* Transfer RNA (tRNA)
tRNA selection pathway, 479–480
tRNA translocation process, 477–479
Tryptic digests, interpretating product ion spectra of, 107–108
Tryptic peptides, amino acid-containing, 597t
Tryptophan, characteristics of, 21t
Tryptophanyl-tRNA synthetase, Rossmann fold in, 127
Turnover number, 335
Turn regions, 115–116
26S proteasomes, 432–433
Twilight zone, 598
Twist angle, 65
Two-dimensional difference in-gel electrophoresis (2D-DIGE), 630
Two-dimensional Fourier transform NMR, 202–203
Two-dimensional gel electrophoresis (2DE), 43–44, 629–631
Two-dimensional polyacrylamide gel electrophoresis (2D-PAGE), 43. *See also* SWISS-2DPAGE Web site; WORLD-2DPAGE index
  uses of, 44
Two-hybrid assay, protein–protein interaction analysis by, 636–637
Two-state transition models, 268–271
Type I topoisomerases, 451–453
Type II restriction enzymes, 57, 59t
Type II topoisomerases, 453
Tyrosine, characteristics of, 21t
Tyrosine kinase-containing receptors, 417–418
Tyrosine kinases, 414
Tyrosine phosphosphorylation, 417

U
Ubiquitin protein degradation system, 430–433
Ubiquitin proteolytic pathway, 431
Ub-protein ligases, 431–432
Ultraviolet (UV) spectra, 185–187. *See also* UV/visible spectroscopic probe
Ultraviolet spectroscopy, 185–190
Uniform Resource Locator (URL), 544
UniProt Archive (UniParc), 601–602
UniProt consortium, 95, 601
UniProt Knowledgebase (UniProtKB), 602
UniProt Reference databases (UniRef), 602
Unitary scoring matrix, 521t
Unit cells, in crystals, 214
Universal tree of life, 698
Untranslated regions (UTRs), 563
  detecting, 578
Unweighed pair group method using arithmetic means (UPGMA), 690
Up-and-down β barrels, 120
Usage directed method, 261
UV/visible spectroscopic probe, 352. *See also* Ultraviolet (UV) spectra

V
Valine, characteristics of, 19t
van der Waals interactions, 5, 6, 255
van der Waals radii, 4, 8
Variable number tandem repeats (VNTRs), 565
Variation, germ-line mechanisms of, 701
Vector alignment search tool (VAST), 610
Vectors, 507–508
Vibrational frequency, 193
Vibrational modes, 193
Viral genomes, 558
Viral reverse transcriptase, 455
Viroids, 387, 391
Virusoids, 387, 391
Visible absorption spectroscopy, 185–190
V-J/V-D-J joining, in gene assembly, 506

W
Watson–Crick base-pairing, 63, 66
  alteration to, 72
Web catalogues, 544–545
Webcutter Web site, 562
Web databases, 545
Web directories, 544–545
WebEntrez, 551. *See also* Entrez entries
WebPHYLIP Web site, 697
Web Primer Web site, 497
Wheatgerm agglutinin (WGA), 312, 315, 317
  NeuNAc binding to, 320
Whole genome shotgun (WGS) approach, 564
WIGS Web site, 562

WORLD-2DPAGE index, 631
World Wide Web (WWW), 543–545
World Wide Web Consortium (W3C), 546
WPDB loader (WPDBL), 607
WPDB Web site, 607

X
Xenobiotics, 399
Xenometabolism, 399
X-ray crystallographic studies, of enzyme mechanisms, 344, 361
X-ray crystal structure, 217
X-ray diffraction patterns, for parent and isomorphous crystals, 216
X-ray diffraction spectroscopy, 214–219

Y
Yahoo, 545
Yeast Proteome Database (YPD), 605

Z
Z-form DNA, 69, 72
Zinc finger motif, 309
Zipper transition model, 270–271
Zone electrophoresis, 41, 43
Z-score, 524
Zwitterions, 21
Zymogens, 374